Department of Computer Science
American International University-Bangladesh

Course Name: INTRODUCTION TO DATA SCIENCE

"Project on Interactive Dashboard using Shiny based on Web Scraping Data"

Section: B

Supervised By:

Dr. Akinul Islam Jony

Associate Professor & Head-In-Charge [Undergraduate Program], Computer Science

Submitted By:

Abdullah,C.M (18-38631-2)

Sumaiya Malik (20-43688-2)

Soily Ghosh Sneha (20-43702-2)

Mubasshar-Ul-Ishraq Tamim (20-43814-2)

Submission Date: May 3, 2023.

Project Title: Interactive Dashboard using Shiny based on Web Scraping Data.

Project Overview:

For this project, we have been assigned to scrap data from webpages, perform pre-processing techniques (Data cleaning, Integration, Data Transformation, Data Reduction, Data Discretization) on them, describe them in the light of descriptive statistics and visualize them using R language then display them in shiny dashboard.

The process of extracting data from a webpage is known as web scraping. This data is gathered and then exported in a way that the user will find more valuable. A spreadsheet or an API, for example. There are many techniques for web scraping. we have used SelectorGadget tool to scraping our data table.

First, we used Outfield Players data for Real Madrid, one of the most successful club of this age, in our project. We obtained Real Madrid's data for the 2022-2023 season from the ESPN website. Then, the datasets were combined. After that, we compared a lot of information, Real Madrid's such as winning record and player performance, and we looked at the dataset. Real-world data is frequently insufficient, unreliable, and loud; thus, it must be cleaned up before being used for the intended purpose. This is often referred to as data pre-processing. Data pre-processing is a data mining technique used to transform unstructured data into something usable and useful. Data Cleansing, Data Integration, Data Transformation, Data Reduction, and Data Discretization are the most crucial steps in the pre-processing of data. Whenever necessary, we pre-processed the data. With the aid of descriptive approaches, we described our data in descriptive analysis. We describe our data in some way and present it in a meaningful way in the descriptive analysis so that it is clear to the reader. We used the Mean, Median, Mode, Range, Variance, Quartile, and Percentile to describe a comparison between various things. Last but not least, we used data visualization to make sense of the data and to help the reader comprehend it. Users may convey insights through graphics far more easily and effectively than through words, and they can also have a bigger impact. Here, we attempted to visually represent practically all aspects of comparison and relationship. Lastly, showed the displayed information in the shiny dashboard.

Project Solution Design:

In order to prepare the dataset for data analysis, we first collected our player lists for Real Madrid from a number of sources. The data was then saved as a CSV file. The next step is to pre-process the data. Data cleaning is the process of going over a raw dataset to look for and get rid of mistakes, duplicates, and unnecessary data. The table contained some missing data, which we filled in with the median after replacing with N/A. Next, we made an effort to manage each piece of noisy data that was present in the dataset. Following data cleaning, steps were made to further clean the data set through data integration, data transformation, data reduction, and data discretization. After completing the data preparation, we focused on applying descriptive statistics to logically simplify our vast amounts of data. In addition, a summary of the dataset's approximations. The metrics Mean, Median, Mode, Range, Variance, Standard Deviation, Quartiles, Percentiles, and Interquartile Ranges were used to collect the data. After finalizing the descriptive statistics, we employed data visualization to visually convey facts and data.

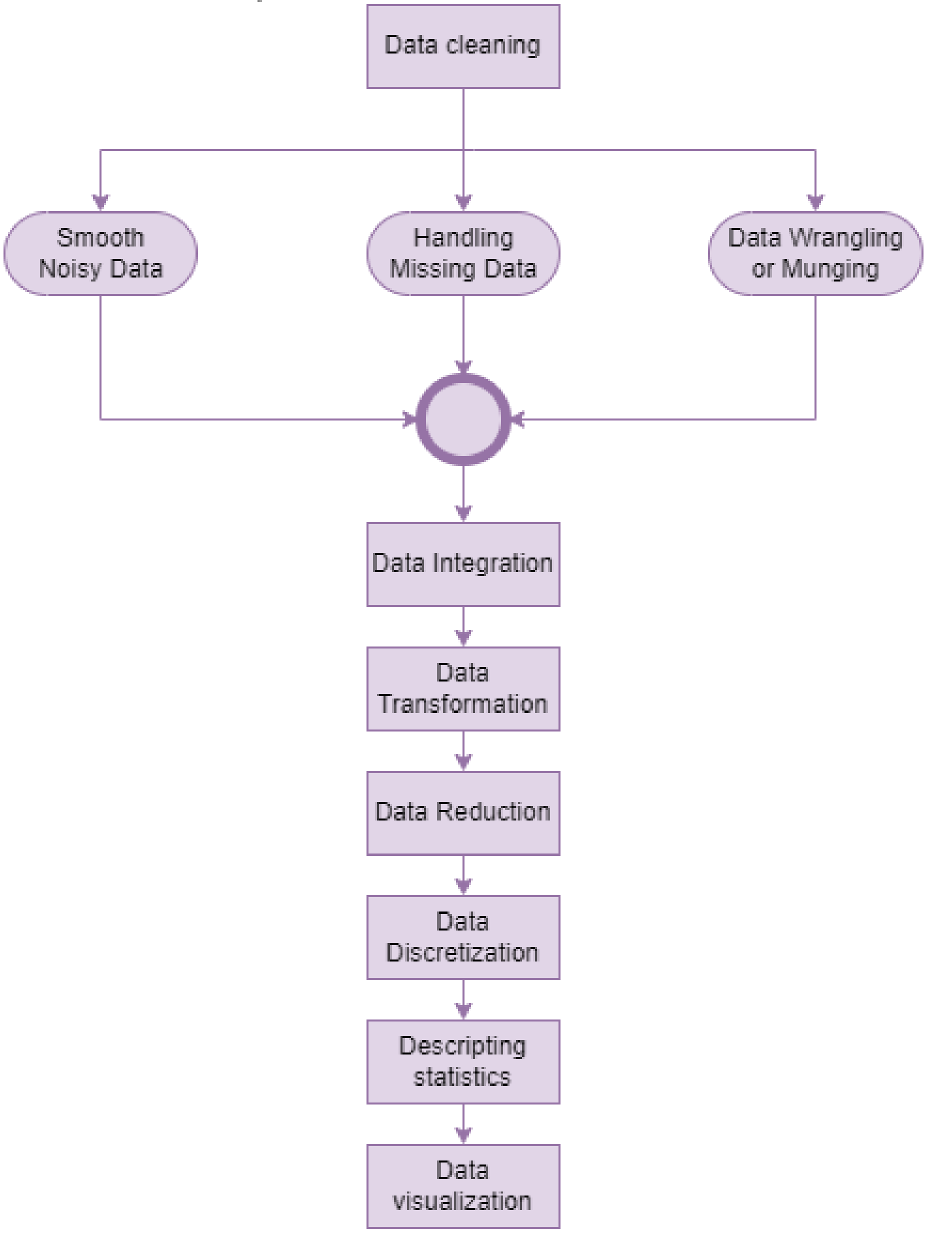
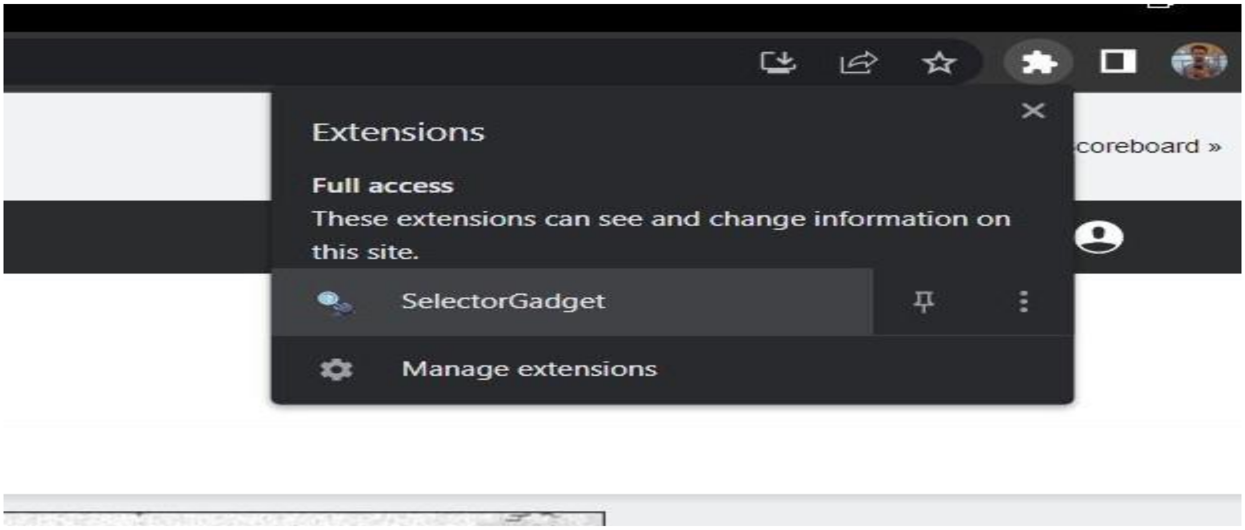


Figure 1: Block Diagram of the project Solution.

Data Collection:

For this project, we start to scrap the data from the website. First, we start to scrap the data from team Real Madrid. In this process, we use a SelectorGadget to simply select data on a website and it will determine its HTML/CSS tags, IDs and classes.



Extensions

Full access

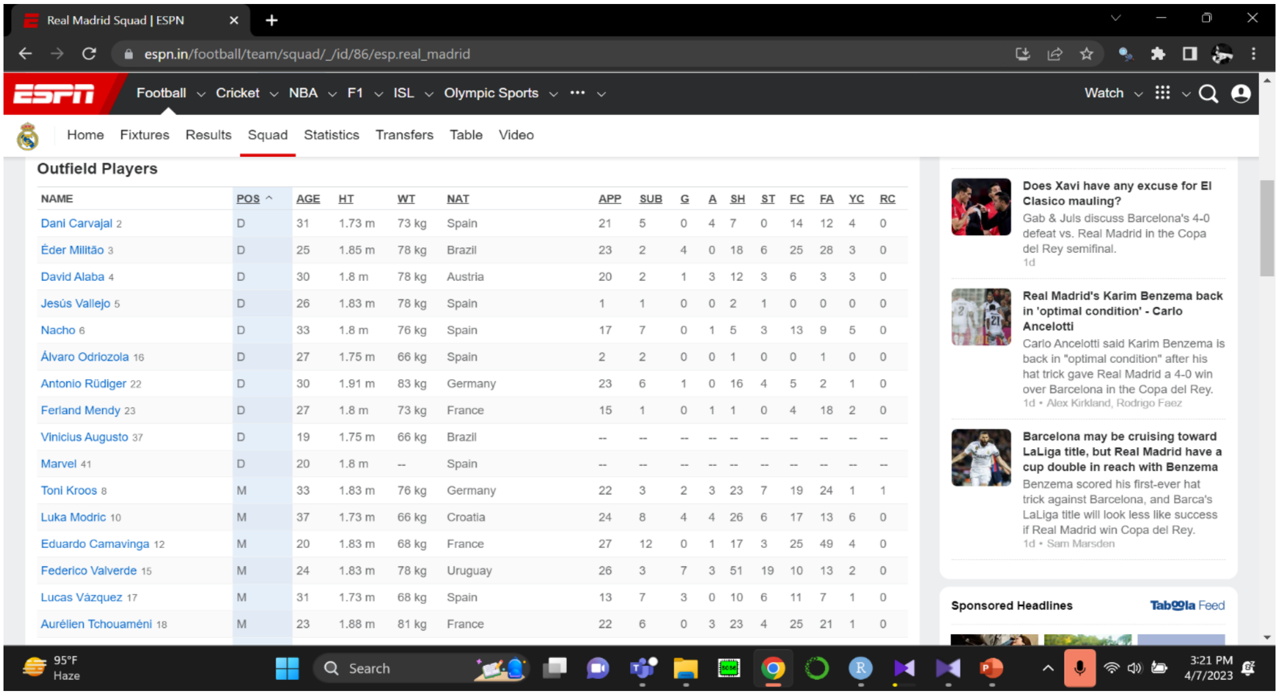
These extensions can see and change information on this site.

SelectorGadget

Manage extensions

Figure 2: Using SelectorGadget Tool for Web Scraping.

Getting Data for Real Madrid:



Outfield Players

NAME	POS	AGE	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC
Dani Carvajal 2	D	31	1.73 m	73 kg	Spain	21	5	0	4	7	0	14	12	4	0
Éder Militão 3	D	25	1.85 m	78 kg	Brazil	23	2	4	0	18	6	25	28	3	0
David Alaba 4	D	30	1.8 m	78 kg	Austria	20	2	1	3	12	3	6	3	3	0
Jesús Vallejo 5	D	26	1.83 m	78 kg	Spain	1	1	0	0	2	1	0	0	0	0
Nacho 6	D	33	1.8 m	76 kg	Spain	17	7	0	1	5	3	13	9	5	0
Álvaro Odriozola 16	D	27	1.75 m	66 kg	Spain	2	2	0	0	1	0	0	1	0	0
Antonio Rüdiger 22	D	30	1.91 m	83 kg	Germany	23	6	1	0	16	4	5	2	1	0
Ferland Mendy 23	D	27	1.8 m	73 kg	France	15	1	0	1	1	0	4	18	2	0
Vinicius Augusto 37	D	19	1.75 m	66 kg	Brazil	--	--	--	--	--	--	--	--	--	--
Marvel 41	D	20	1.8 m	--	Spain	--	--	--	--	--	--	--	--	--	--
Toni Kroos 8	M	33	1.83 m	76 kg	Germany	22	3	2	3	23	7	19	24	1	1
Luka Modric 10	M	37	1.73 m	66 kg	Croatia	24	8	4	4	26	6	17	13	6	0
Eduardo Camavinga 12	M	20	1.83 m	68 kg	France	27	12	0	1	17	3	25	49	4	0
Federico Valverde 15	M	24	1.83 m	78 kg	Uruguay	26	3	7	3	51	19	10	13	2	0
Lucas Vázquez 17	M	31	1.73 m	68 kg	Spain	13	7	3	0	10	6	11	7	1	0
Aurélien Tchouaméni 18	M	23	1.88 m	81 kg	France	22	6	0	3	23	4	25	21	1	0

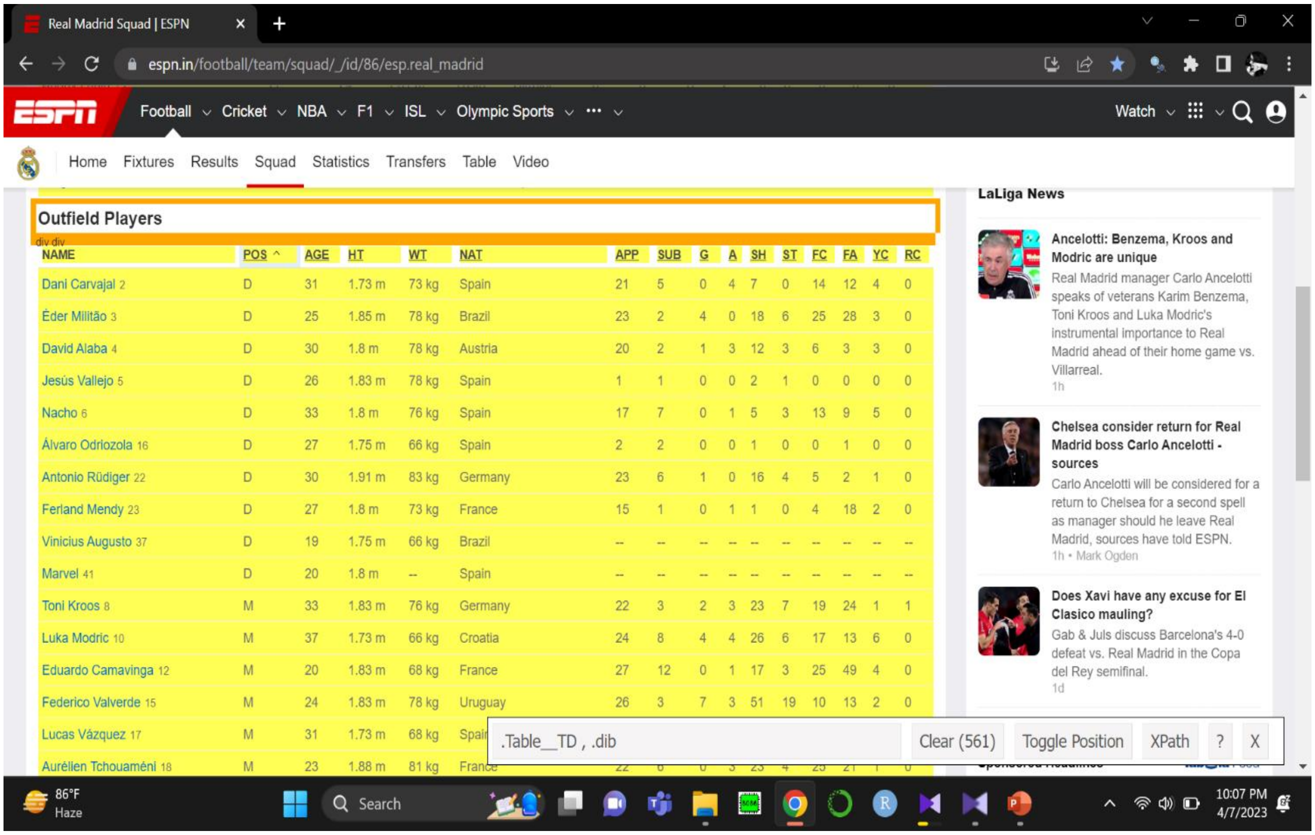
Does Xavi have any excuse for El Clasico mauling?
Gab & Juls discuss Barcelona's 4-0 defeat vs. Real Madrid in the Copa del Rey semifinal.
1d

Real Madrid's Karim Benzema back in 'optimal condition' - Carlo Ancelotti
Carlo Ancelotti said Karim Benzema is back in "optimal condition" after his hat trick gave Real Madrid a 4-0 win over Barcelona in the Copa del Rey.
1d • Alex Kirkland, Rodrigo Faez

Barcelona may be cruising toward LaLiga title, but Real Madrid have a cup double in reach with Benzema
Benzema scored his first-ever hat trick against Barcelona, and Barca's LaLiga title will look less like success if Real Madrid win Copa del Rey.
1d • Sam Marsden

Sponsored Headlines Taboola Feed

Figure 3: Real Madrid’s data for the 2022-2023 season from the ESPN website.



Outfield Players

NAME	POS ^	AGE	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC
Dani Carvajal 2	D	31	1.73 m	73 kg	Spain	21	5	0	4	7	0	14	12	4	0
Éder Militão 3	D	25	1.85 m	78 kg	Brazil	23	2	4	0	18	6	25	28	3	0
David Alaba 4	D	30	1.8 m	78 kg	Austria	20	2	1	3	12	3	6	3	3	0
Jesús Vallejo 5	D	26	1.83 m	78 kg	Spain	1	1	0	0	2	1	0	0	0	0
Nacho 6	D	33	1.8 m	76 kg	Spain	17	7	0	1	5	3	13	9	5	0
Álvaro Odriozola 16	D	27	1.75 m	66 kg	Spain	2	2	0	0	1	0	0	1	0	0
Antonio Rüdiger 22	D	30	1.91 m	83 kg	Germany	23	6	1	0	16	4	5	2	1	0
Ferland Mendy 23	D	27	1.8 m	73 kg	France	15	1	0	1	1	0	4	18	2	0
Vinicius Augusto 37	D	19	1.75 m	66 kg	Brazil	--	--	--	--	--	--	--	--	--	--
Marvel 41	D	20	1.8 m	--	Spain	--	--	--	--	--	--	--	--	--	--
Toni Kroos 8	M	33	1.83 m	76 kg	Germany	22	3	2	3	23	7	19	24	1	1
Luka Modric 10	M	37	1.73 m	66 kg	Croatia	24	8	4	4	26	6	17	13	6	0
Eduardo Camavinga 12	M	20	1.83 m	68 kg	France	27	12	0	1	17	3	25	49	4	0
Federico Valverde 15	M	24	1.83 m	78 kg	Uruguay	26	3	7	3	51	19	10	13	2	0
Lucas Vázquez 17	M	31	1.73 m	68 kg	Spai										
Aurélien Tchouaméni 18	M	23	1.88 m	81 kg	France	22	6	0	3	23	4	25	21	1	0

LaLiga News

Ancelotti: Benzema, Kroos and Modric are unique
Real Madrid manager Carlo Ancelotti speaks of veterans Karim Benzema, Toni Kroos and Luka Modric's instrumental importance to Real Madrid ahead of their home game vs. Villarreal.
1h

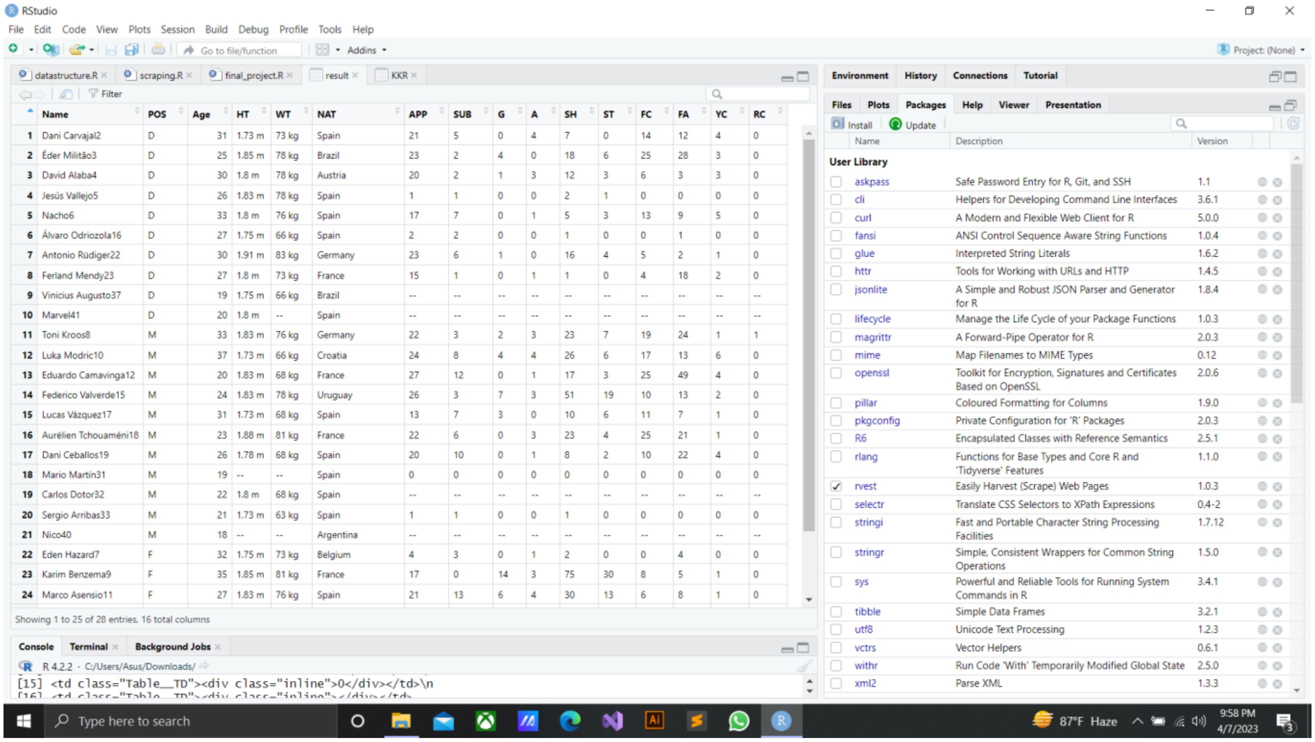
Chelsea consider return for Real Madrid boss Carlo Ancelotti - sources
Carlo Ancelotti will be considered for a return to Chelsea for a second spell as manager should he leave Real Madrid, sources have told ESPN.
1h • Mark Ogden

Does Xavi have any excuse for El Clasico mauling?
Gab & Juls discuss Barcelona's 4-0 defeat vs. Real Madrid in the Copa del Rey semifinal.
1d

Figure 4: Real Madrid's data while Web Scraping Using SelectorGadget Tool.

Code:

```
library(rvest)  #rvest helps you scrape

football = read_html("https://www.espn.in/football/team/squad/_/id/86/esp.real_madrid")

ft = html_nodes(football, css=".Table__TD")

ft

result <-data.frame(html_table(football, header = TRUE)[[2]])

View(result)

write.csv(result,"D:\\result.csv")
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
datastructure.R
scraping.R
final_project.R
result
KKR
Filter
Environment History Connections Tutorial
Files Plots Packages Help Viewer Presentation
Install Update
Project: (None)
Console Terminal Background Jobs
Type here to search
87°F Haze
9:58 PM
4/7/2023

	Name	POS	Age	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC
1	Dani Carvajal2	D	31	1.73 m	73 kg	Spain	21	5	0	4	7	0	14	12	4	0
2	Éder Militão3	D	25	1.85 m	78 kg	Brazil	23	2	4	0	18	6	25	28	3	0
3	David Alaba4	D	30	1.8 m	78 kg	Austria	20	2	1	3	12	3	6	3	3	0
4	Jesús Vallejo5	D	26	1.83 m	78 kg	Spain	1	1	0	0	2	1	0	0	0	0
5	Nacho6	D	33	1.8 m	76 kg	Spain	17	7	0	1	5	3	13	9	5	0
6	Álvaro Odriozola16	D	27	1.75 m	66 kg	Spain	2	2	0	0	1	0	0	1	0	0
7	Antonio Rüdiger22	D	30	1.91 m	83 kg	Germany	23	6	1	0	16	4	5	2	1	0
8	Ferland Mendy23	D	27	1.8 m	73 kg	France	15	1	0	1	1	0	4	18	2	0
9	Vinicius Augusto37	D	19	1.75 m	66 kg	Brazil	--	--	--	--	--	--	--	--	--	--
10	Marvel41	D	20	1.8 m	--	Spain	--	--	--	--	--	--	--	--	--	--
11	Toni Kroos8	M	33	1.83 m	76 kg	Germany	22	3	2	3	23	7	19	24	1	1
12	Luka Modric10	M	37	1.73 m	66 kg	Croatia	24	8	4	4	26	6	17	13	6	0
13	Eduardo Camavinga12	M	20	1.83 m	68 kg	France	27	12	0	1	17	3	25	49	4	0
14	Federico Valverde15	M	24	1.83 m	78 kg	Uruguay	26	3	7	3	51	19	10	13	2	0
15	Lucas Vázquez17	M	31	1.73 m	68 kg	Spain	13	7	3	0	10	6	11	7	1	0
16	Aurélien Tchouaméni18	M	23	1.88 m	81 kg	France	22	6	0	3	23	4	25	21	1	0
17	Dani Ceballos19	M	26	1.78 m	68 kg	Spain	20	10	0	1	8	2	10	22	4	0
18	Mario Martín31	M	19	--	--	Spain	0	0	0	0	0	0	0	0	0	0
19	Carlos Dotor32	M	22	1.8 m	68 kg	Spain	--	--	--	--	--	--	--	--	--	--
20	Sergio Arribas33	M	21	1.73 m	63 kg	Spain	1	1	0	0	1	0	0	0	0	0
21	Nico40	M	18	--	--	Argentina	--	--	--	--	--	--	--	--	--	--
22	Eden Hazard7	F	32	1.75 m	73 kg	Belgium	4	3	0	1	2	0	0	4	0	0
23	Karim Benzema9	F	35	1.85 m	81 kg	France	17	0	14	3	75	30	8	5	1	0
24	Marco Asensio11	F	27	1.83 m	76 kg	Spain	21	13	6	4	30	13	6	8	1	0

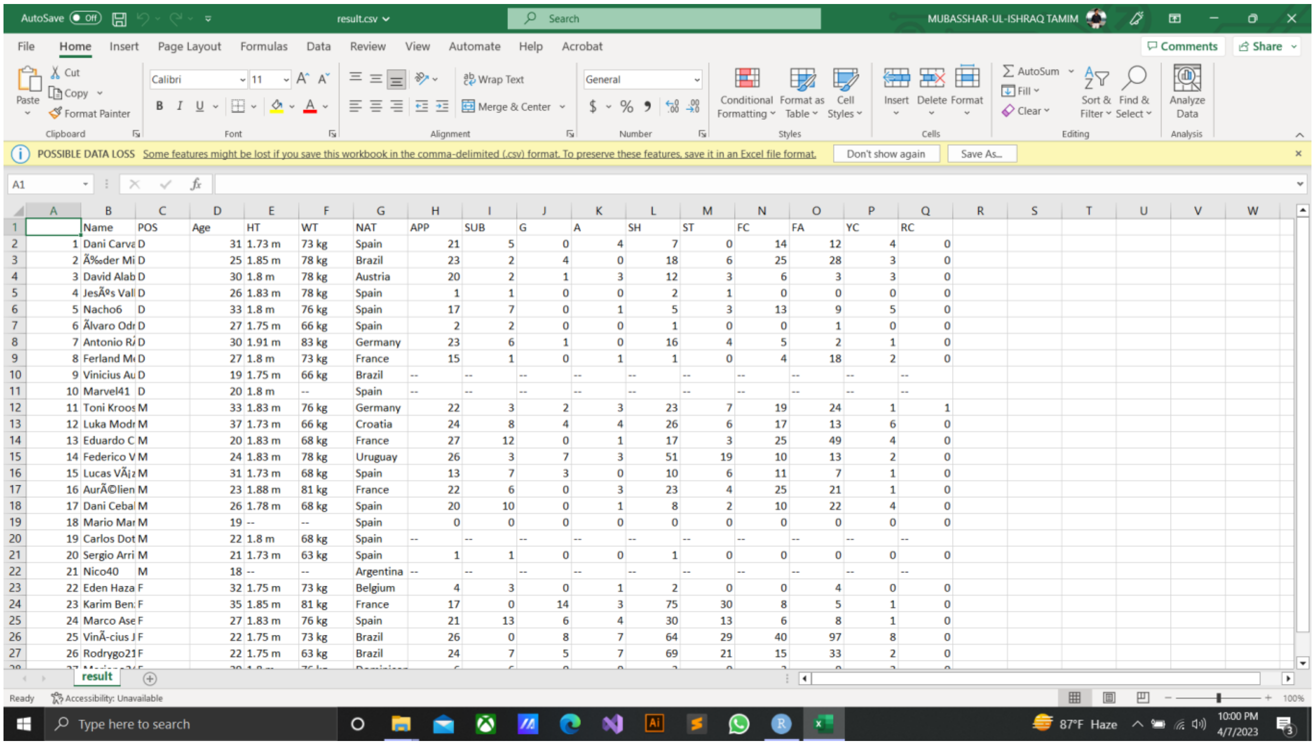
Showing 1 to 25 of 28 entries, 16 total columns

```
R 4.2.2 · C:/Users/Asus/Downloads/
[15] <td class="Table__TD"><div class="inline">0</div></td>\n
```

User Library

	Name	Description	Version
□	askpass	Safe Password Entry for R, Git, and SSH	1.1
□	cli	Helpers for Developing Command Line Interfaces	3.6.1
□	curl	A Modern and Flexible Web Client for R	5.0.0
□	fansi	ANSI Control Sequence Aware String Functions	1.0.4
□	glue	Interpreted String Literals	1.6.2
□	httr	Tools for Working with URLs and HTTP	1.4.5
□	jsonlite	A Simple and Robust JSON Parser and Generator for R	1.8.4
□	lifecycle	Manage the Life Cycle of your Package Functions	1.0.3
□	magrittr	A Forward-Pipe Operator for R	2.0.3
□	mime	Map Filenames to MIME Types	0.12
□	openssl	Toolkit for Encryption, Signatures and Certificates Based on OpenSSL	2.0.6
□	pillar	Coloured Formatting for Columns	1.9.0
□	pkgconfig	Private Configuration for 'R' Packages	2.0.3
□	R6	Encapsulated Classes with Reference Semantics	2.5.1
□	rlang	Functions for Base Types and Core R and 'Tidyverse' Features	1.1.0
☑	rvest	Easily Harvest (Scrape) Web Pages	1.0.3
□	selectr	Translate CSS Selectors to XPath Expressions	0.4-2
□	stringi	Fast and Portable Character String Processing Facilities	1.7.12
□	stringr	Simple, Consistent Wrappers for Common String Operations	1.5.0
□	sys	Powerful and Reliable Tools for Running System Commands in R	3.4.1
□	tibble	Simple Data Frames	3.2.1
□	utf8	Unicode Text Processing	1.2.3
□	vctrs	Vector Helpers	0.6.1
□	withr	Run Code 'With' Temporarily Modified Global State	2.5.0
□	xml2	Parse XML	1.3.3

Figure 5: Real Madrid's data Output in RStudio.



POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		Name	POS	Age	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC
2	1	Dani Carva	D	31	1.73 m	73 kg	Spain	21	5	0	4	7	0	14	12	4	0
3	2	Ã‰der Mi	D	25	1.85 m	78 kg	Brazil	23	2	4	0	18	6	25	28	3	0
4	3	David Alab	D	30	1.8 m	78 kg	Austria	20	2	1	3	12	3	6	3	3	0
5	4	JesÃºs Vall	D	26	1.83 m	78 kg	Spain	1	1	0	0	2	1	0	0	0	0
6	5	Nacho6	D	33	1.8 m	76 kg	Spain	17	7	0	1	5	3	13	9	5	0
7	6	Ãlvaro Odr	D	27	1.75 m	66 kg	Spain	2	2	0	0	1	0	0	1	0	0
8	7	Antonio RÃ	D	30	1.91 m	83 kg	Germany	23	6	1	0	16	4	5	2	1	0
9	8	Ferland M	D	27	1.8 m	73 kg	France	15	1	0	1	1	0	4	18	2	0
10	9	Vinicius Au	D	19	1.75 m	66 kg	Brazil	--	--	--	--	--	--	--	--	--	--
11	10	Marvel41	D	20	1.8 m	--	Spain	--	--	--	--	--	--	--	--	--	--
12	11	Toni Kroos	M	33	1.83 m	76 kg	Germany	22	3	2	3	23	7	19	24	1	1
13	12	Luka Modr	M	37	1.73 m	66 kg	Croatia	24	8	4	4	26	6	17	13	6	0
14	13	Eduardo C	M	20	1.83 m	68 kg	France	27	12	0	1	17	3	25	49	4	0
15	14	Federico V	M	24	1.83 m	78 kg	Uruguay	26	3	7	3	51	19	10	13	2	0
16	15	Lucas VÃ¡z	M	31	1.73 m	68 kg	Spain	13	7	3	0	10	6	11	7	1	0
17	16	AurÃ©lien	M	23	1.88 m	81 kg	France	22	6	0	3	23	4	25	21	1	0
18	17	Dani Cebal	M	26	1.78 m	68 kg	Spain	20	10	0	1	8	2	10	22	4	0
19	18	Mario Mar	M	19	--	--	Spain	0	0	0	0	0	0	0	0	0	0
20	19	Carlos Dot	M	22	1.8 m	68 kg	Spain	--	--	--	--	--	--	--	--	--	--
21	20	Sergio Arri	M	21	1.73 m	63 kg	Spain	1	1	0	0	1	0	0	0	0	0
22	21	Nico40	M	18	--	--	Argentina	--	--	--	--	--	--	--	--	--	--
23	22	Eden Haza	F	32	1.75 m	73 kg	Belgium	4	3	0	1	2	0	0	4	0	0
24	23	Karim Ben	F	35	1.85 m	81 kg	France	17	0	14	3	75	30	8	5	1	0
25	24	Marco Ase	F	27	1.83 m	76 kg	Spain	21	13	6	4	30	13	6	8	1	0
26	25	VinÃcius J	F	22	1.75 m	73 kg	Brazil	26	0	8	7	64	29	40	97	8	0
27	26	Rodrygo21	F	22	1.75 m	63 kg	Brazil	24	7	5	7	69	21	15	33	2	0

Figure 6: Real Madrid's data saved in CSV File.

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 assigned. To do so we write a code that will show us the row which contains the missing value.

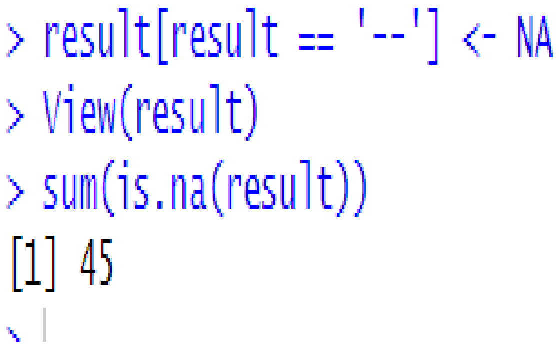
Code:

#Missing data replace by N/A

```
result[result == '--'] <- NA
View(result)
```

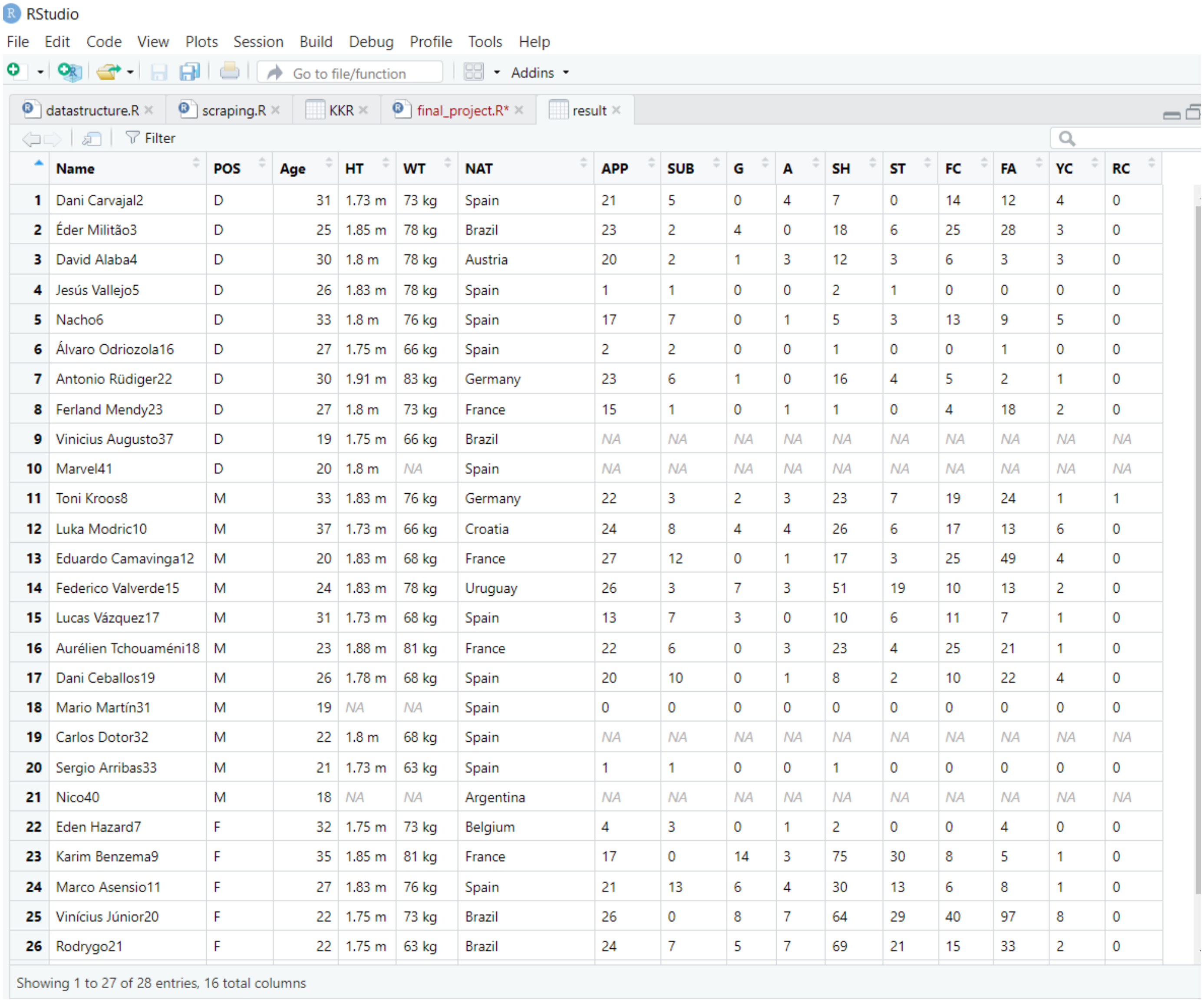
#Number of missing data

```
sum(is.na(result))
```



```
> result[result == '--'] <- NA
> View(result)
> sum(is.na(result))
[1] 45
>
```

Figure 7: Code of missing data replace by N/A.



	Name	POS	Age	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC
1	Dani Carvajal2	D	31	1.73 m	73 kg	Spain	21	5	0	4	7	0	14	12	4	0
2	Éder Militão3	D	25	1.85 m	78 kg	Brazil	23	2	4	0	18	6	25	28	3	0
3	David Alaba4	D	30	1.8 m	78 kg	Austria	20	2	1	3	12	3	6	3	3	0
4	Jesús Vallejo5	D	26	1.83 m	78 kg	Spain	1	1	0	0	2	1	0	0	0	0
5	Nacho6	D	33	1.8 m	76 kg	Spain	17	7	0	1	5	3	13	9	5	0
6	Álvaro Odriozola16	D	27	1.75 m	66 kg	Spain	2	2	0	0	1	0	0	1	0	0
7	Antonio Rüdiger22	D	30	1.91 m	83 kg	Germany	23	6	1	0	16	4	5	2	1	0
8	Ferland Mendy23	D	27	1.8 m	73 kg	France	15	1	0	1	1	0	4	18	2	0
9	Vinicius Augusto37	D	19	1.75 m	66 kg	Brazil	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	Marvel41	D	20	1.8 m	NA	Spain	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	Toni Kroos8	M	33	1.83 m	76 kg	Germany	22	3	2	3	23	7	19	24	1	1
12	Luka Modric10	M	37	1.73 m	66 kg	Croatia	24	8	4	4	26	6	17	13	6	0
13	Eduardo Camavinga12	M	20	1.83 m	68 kg	France	27	12	0	1	17	3	25	49	4	0
14	Federico Valverde15	M	24	1.83 m	78 kg	Uruguay	26	3	7	3	51	19	10	13	2	0
15	Lucas Vázquez17	M	31	1.73 m	68 kg	Spain	13	7	3	0	10	6	11	7	1	0
16	Aurélien Tchouaméni18	M	23	1.88 m	81 kg	France	22	6	0	3	23	4	25	21	1	0
17	Dani Ceballos19	M	26	1.78 m	68 kg	Spain	20	10	0	1	8	2	10	22	4	0
18	Mario Martín31	M	19	NA	NA	Spain	0	0	0	0	0	0	0	0	0	0
19	Carlos Dotor32	M	22	1.8 m	68 kg	Spain	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
20	Sergio Arribas33	M	21	1.73 m	63 kg	Spain	1	1	0	0	1	0	0	0	0	0
21	Nico40	M	18	NA	NA	Argentina	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
22	Eden Hazard7	F	32	1.75 m	73 kg	Belgium	4	3	0	1	2	0	0	4	0	0
23	Karim Benzema9	F	35	1.85 m	81 kg	France	17	0	14	3	75	30	8	5	1	0
24	Marco Asensio11	F	27	1.83 m	76 kg	Spain	21	13	6	4	30	13	6	8	1	0
25	Vinícius Júnior20	F	22	1.75 m	73 kg	Brazil	26	0	8	7	64	29	40	97	8	0
26	Rodrygo21	F	22	1.75 m	63 kg	Brazil	24	7	5	7	69	21	15	33	2	0

Showing 1 to 27 of 28 entries, 16 total columns

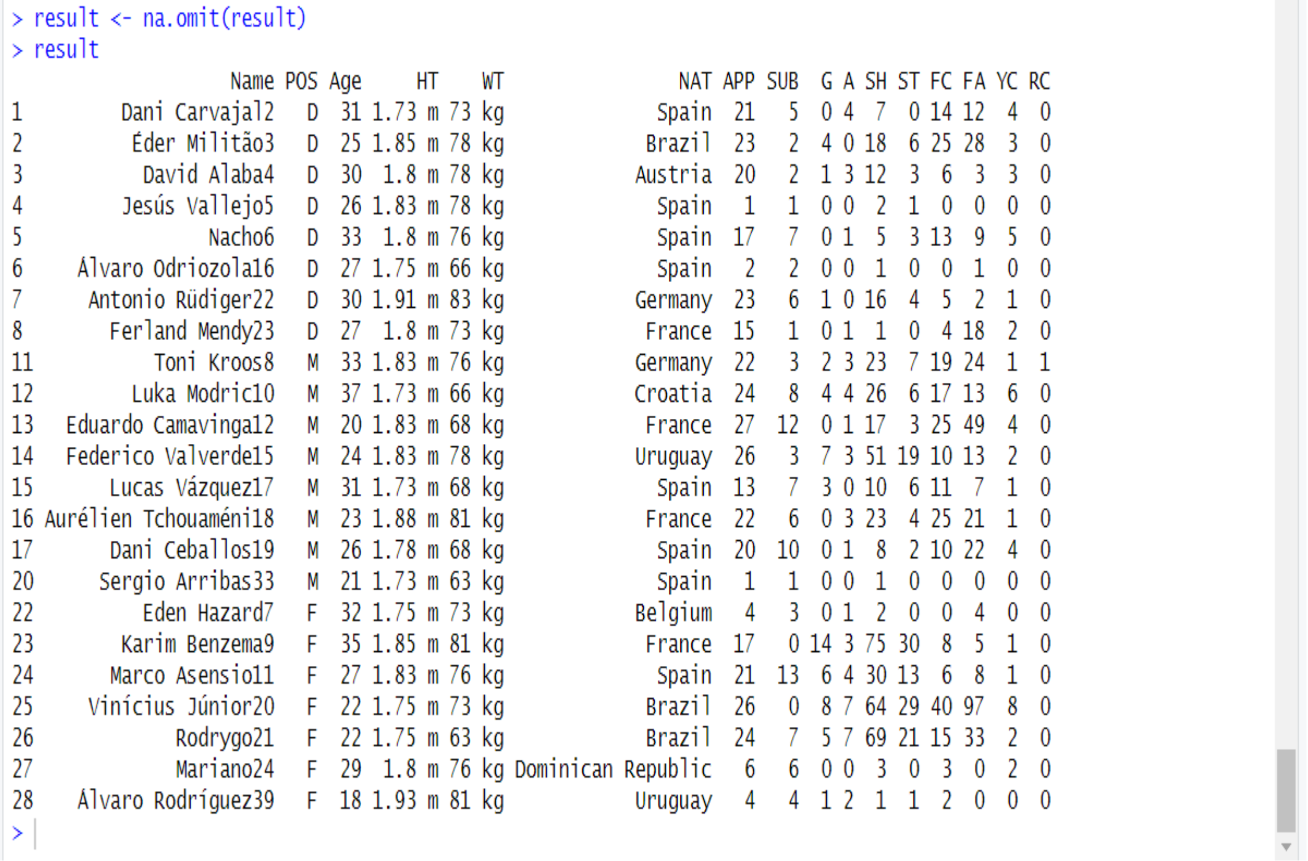
Figure 8: Missing data replace by N/A.

The handling method must now be carried out after we have identified the missing data. These are significant player statistics regarding the performance and season as a whole, as we can see. As a result, since missing data cannot be filled in by any method or assumption, all players with missing data must be eliminated from the data set.

<u>Code:</u>

<u>#Omit the missing data</u>

```
result <- na.omit(result)
result
```



```
> result <- na.omit(result)
> result
                    Name POS Age     HT    WT                NAT APP SUB  G A SH ST FC FA YC RC
1         Dani Carvajal2   D  31 1.73 m 73 kg              Spain  21   5  0 4  7  0 14 12  4  0
2          Éder Militão3   D  25 1.85 m 78 kg             Brazil  23   2  4 0 18  6 25 28  3  0
3           David Alaba4   D  30  1.8 m 78 kg            Austria  20   2  1 3 12  3  6  3  3  0
4         Jesús Vallejo5   D  26 1.83 m 78 kg              Spain   1   1  0 0  2  1  0  0  0  0
5                 Nacho6   D  33  1.8 m 76 kg              Spain  17   7  0 1  5  3 13  9  5  0
6     Álvaro Odriozola16   D  27 1.75 m 66 kg              Spain   2   2  0 0  1  0  0  1  0  0
7      Antonio Rüdiger22   D  30 1.91 m 83 kg            Germany  23   6  1 0 16  4  5  2  1  0
8        Ferland Mendy23   D  27  1.8 m 73 kg             France  15   1  0 1  1  0  4 18  2  0
11           Toni Kroos8   M  33 1.83 m 76 kg            Germany  22   3  2 3 23  7 19 24  1  1
12         Luka Modric10   M  37 1.73 m 66 kg            Croatia  24   8  4 4 26  6 17 13  6  0
13   Eduardo Camavinga12   M  20 1.83 m 68 kg             France  27  12  0 1 17  3 25 49  4  0
14   Federico Valverde15   M  24 1.83 m 78 kg            Uruguay  26   3  7 3 51 19 10 13  2  0
15       Lucas Vázquez17   M  31 1.73 m 68 kg              Spain  13   7  3 0 10  6 11  7  1  0
16 Aurélien Tchouaméni18   M  23 1.88 m 81 kg             France  22   6  0 3 23  4 25 21  1  0
17       Dani Ceballos19   M  26 1.78 m 68 kg              Spain  20  10  0 1  8  2 10 22  4  0
20      Sergio Arribas33   M  21 1.73 m 63 kg              Spain   1   1  0 0  1  0  0  0  0  0
22          Eden Hazard7   F  32 1.75 m 73 kg            Belgium   4   3  0 1  2  0  0  4  0  0
23        Karim Benzema9   F  35 1.85 m 81 kg             France  17   0 14 3 75 30  8  5  1  0
24       Marco Asensio11   F  27 1.83 m 76 kg              Spain  21  13  6 4 30 13  6  8  1  0
25     Vinícius Júnior20   F  22 1.75 m 73 kg             Brazil  26   0  8 7 64 29 40 97  8  0
26             Rodrygo21   F  22 1.75 m 63 kg             Brazil  24   7  5 7 69 21 15 33  2  0
27             Mariano24   F  29  1.8 m 76 kg Dominican Republic   6   6  0 0  3  0  3  0  2  0
28    Álvaro Rodríguez39   F  18 1.93 m 81 kg            Uruguay   4   4  1 2  1  1  2  0  0  0
> 
```

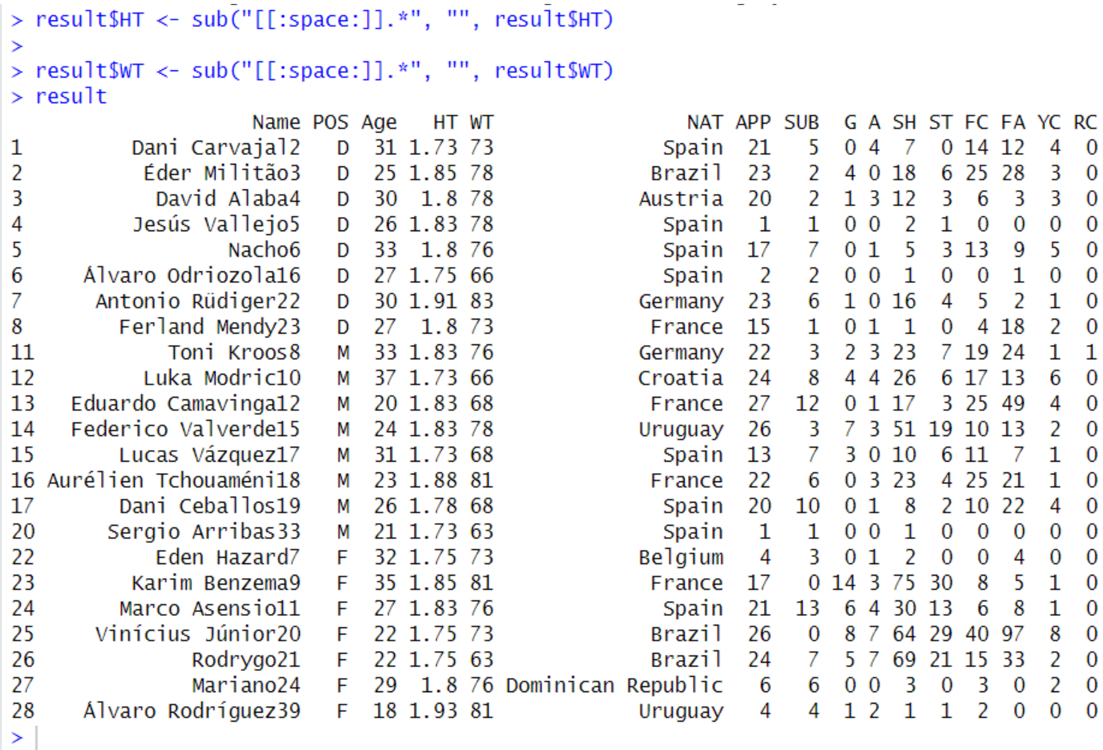
Figure 9: Omit the missing data.

- **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:

# Removing kg and m from wt, ht column

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

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

result
```



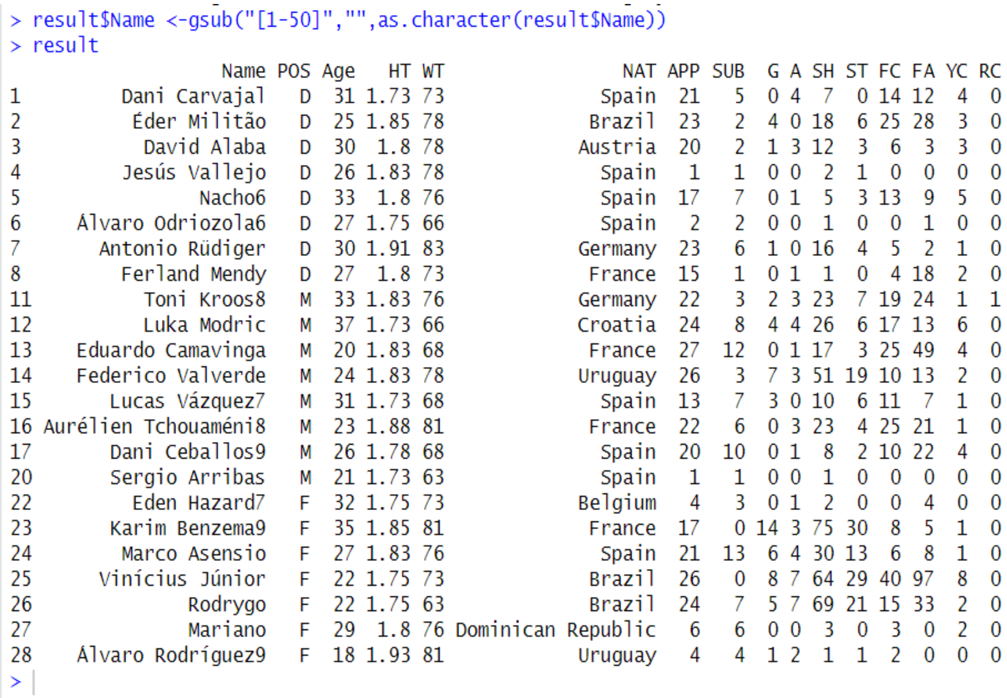
```
> result$HT <- sub("[[:space:]].*", "", result$HT)
>
> result$WT <- sub("[[:space:]].*", "", result$WT)
> result
                    Name POS Age   HT WT                NAT APP SUB  G A SH ST FC FA YC RC
1          Dani Carvajal2   D  31 1.73 73              Spain  21   5  0 4  7  0 14 12  4  0
2           Éder Militão3   D  25 1.85 78             Brazil  23   2  4 0 18  6 25 28  3  0
3            David Alaba4   D  30  1.8 78            Austria  20   2  1 3 12  3  6  3  3  0
4          Jesús Vallejo5   D  26 1.83 78              Spain   1   1  0 0  2  1  0  0  0  0
5                  Nacho6   D  33  1.8 76              Spain  17   7  0 1  5  3 13  9  5  0
6       Álvaro Odriozola16  D  27 1.75 66              Spain   2   2  0 0  1  0  0  1  0  0
7       Antonio Rüdiger22   D  30 1.91 83            Germany  23   6  1 0 16  4  5  2  1  0
8         Ferland Mendy23   D  27  1.8 73             France  15   1  0 1  1  0  4 18  2  0
11            Toni Kroos8   M  33 1.83 76            Germany  22   3  2 3 23  7 19 24  1  1
12          Luka Modric10   M  37 1.73 66            Croatia  24   8  4 4 26  6 17 13  6  0
13     Eduardo Camavinga12  M  20 1.83 68             France  27  12  0 1 17  3 25 49  4  0
14     Federico Valverde15  M  24 1.83 78            Uruguay  26   3  7 3 51 19 10 13  2  0
15         Lucas Vázquez17  M  31 1.73 68              Spain  13   7  3 0 10  6 11  7  1  0
16 Aurélien Tchouaméni18    M  23 1.88 81             France  22   6  0 3 23  4 25 21  1  0
17         Dani Ceballos19  M  26 1.78 68              Spain  20  10  0 1  8  2 10 22  4  0
20        Sergio Arribas33  M  21 1.73 63              Spain   1   1  0 0  1  0  0  0  0  0
22           Eden Hazard7   F  32 1.75 73            Belgium   4   3  0 1  2  0  0  4  0  0
23          Karim Benzema9  F  35 1.85 81             France  17   0 14 3 75 30  8  5  1  0
24        Marco Asensio11   F  27 1.83 76              Spain  21  13  6 4 30 13  6  8  1  0
25       Vinícius Júnior20  F  22 1.75 73             Brazil  26   0  8 7 64 29 40 97  8  0
26              Rodrygo21   F  22 1.75 63             Brazil  24   7  5 7 69 21 15 33  2  0
27              Mariano24   F  29  1.8 76 Dominican Republic   6   6  0 0  3  0  3  0  2  0
28      Álvaro Rodríguez39  F  18 1.93 81            Uruguay   4   4  1 2  1  1  2  0  0  0
>
```

Figure 10: Removing kg and m from wt, ht column.

#Removing numbers from player name

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

result



```
> result$Name <-gsub("[1-50]","",as.character(result$Name))
> result
                   Name POS Age   HT WT                NAT APP SUB  G A SH ST FC FA YC RC
1         Dani Carvajal   D  31 1.73 73              Spain  21   5  0 4  7  0 14 12  4  0
2          Éder Militão   D  25 1.85 78             Brazil  23   2  4 0 18  6 25 28  3  0
3           David Alaba   D  30  1.8 78            Austria  20   2  1 3 12  3  6  3  3  0
4         Jesús Vallejo   D  26 1.83 78              Spain   1   1  0 0  2  1  0  0  0  0
5                Nacho6   D  33  1.8 76              Spain  17   7  0 1  5  3 13  9  5  0
6      Álvaro Odriozola6  D  27 1.75 66              Spain   2   2  0 0  1  0  0  1  0  0
7       Antonio Rüdiger   D  30 1.91 83            Germany  23   6  1 0 16  4  5  2  1  0
8         Ferland Mendy   D  27  1.8 73             France  15   1  0 1  1  0  4 18  2  0
11          Toni Kroos8   M  33 1.83 76            Germany  22   3  2 3 23  7 19 24  1  1
12          Luka Modric   M  37 1.73 66            Croatia  24   8  4 4 26  6 17 13  6  0
13    Eduardo Camavinga   M  20 1.83 68             France  27  12  0 1 17  3 25 49  4  0
14    Federico Valverde   M  24 1.83 78            Uruguay  26   3  7 3 51 19 10 13  2  0
15       Lucas Vázquez7   M  31 1.73 68              Spain  13   7  3 0 10  6 11  7  1  0
16 Aurélien Tchouaméni8   M  23 1.88 81             France  22   6  0 3 23  4 25 21  1  0
17       Dani Ceballos9   M  26 1.78 68              Spain  20  10  0 1  8  2 10 22  4  0
20       Sergio Arribas   M  21 1.73 63              Spain   1   1  0 0  1  0  0  0  0  0
22         Eden Hazard7   F  32 1.75 73            Belgium   4   3  0 1  2  0  0  4  0  0
23       Karim Benzema9   F  35 1.85 81             France  17   0 14 3 75 30  8  5  1  0
24        Marco Asensio   F  27 1.83 76              Spain  21  13  6 4 30 13  6  8  1  0
25      Vinícius Júnior   F  22 1.75 73             Brazil  26   0  8 7 64 29 40 97  8  0
26              Rodrygo   F  22 1.75 63             Brazil  24   7  5 7 69 21 15 33  2  0
27              Mariano   F  29  1.8 76 Dominican Republic   6   6  0 0  3  0  3  0  2  0
28    Álvaro Rodríguez9   F  18 1.93 81            Uruguay   4   4  1 2  1  1  2  0  0  0
>
```

Figure 11: Removing numbers from player name.

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

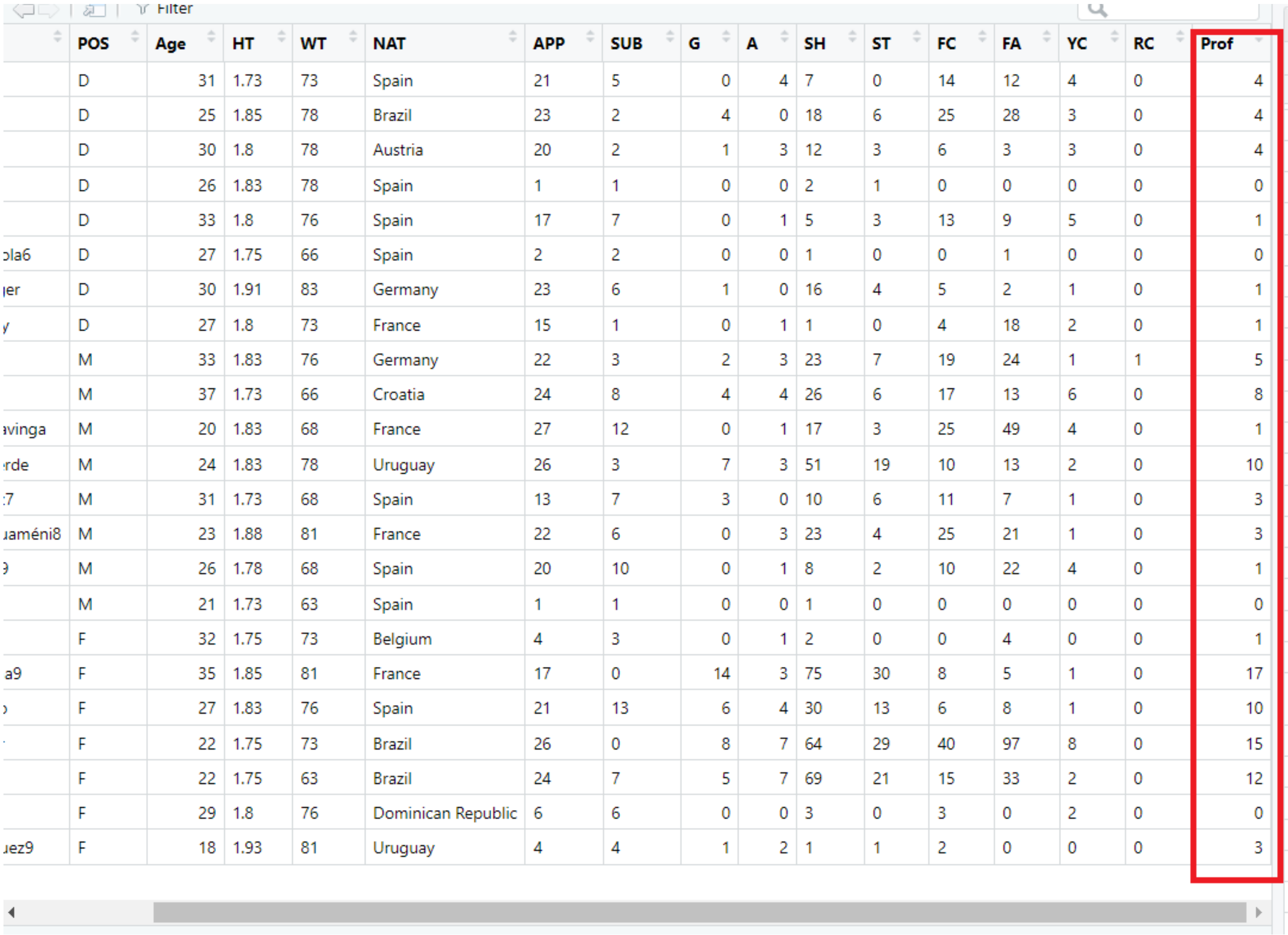
2. **Data Integration:** For the purpose of better analysis, we need to integrate these two data into one complete dataset.

A new column named Performance which is the sum of Goal and Assist,

Code:

#Adding new column

```
result$G <- as.numeric(result$G)

result$A <- as.numeric(result$A)

result$Prof <- result$G + result$A

view(result)
```



	POS	Age	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC	Prof
	D	31	1.73	73	Spain	21	5	0	4	7	0	14	12	4	0	4
	D	25	1.85	78	Brazil	23	2	4	0	18	6	25	28	3	0	4
	D	30	1.8	78	Austria	20	2	1	3	12	3	6	3	3	0	4
	D	26	1.83	78	Spain	1	1	0	0	2	1	0	0	0	0	0
	D	33	1.8	76	Spain	17	7	0	1	5	3	13	9	5	0	1
ɔla6	D	27	1.75	66	Spain	2	2	0	0	1	0	0	1	0	0	0
ıer	D	30	1.91	83	Germany	23	6	1	0	16	4	5	2	1	0	1
y	D	27	1.8	73	France	15	1	0	1	1	0	4	18	2	0	1
	M	33	1.83	76	Germany	22	3	2	3	23	7	19	24	1	1	5
	M	37	1.73	66	Croatia	24	8	4	4	26	6	17	13	6	0	8
ıvinga	M	20	1.83	68	France	27	12	0	1	17	3	25	49	4	0	1
ırde	M	24	1.83	78	Uruguay	26	3	7	3	51	19	10	13	2	0	10
:7	M	31	1.73	68	Spain	13	7	3	0	10	6	11	7	1	0	3
ıaméni8	M	23	1.88	81	France	22	6	0	3	23	4	25	21	1	0	3
ɟ	M	26	1.78	68	Spain	20	10	0	1	8	2	10	22	4	0	1
	M	21	1.73	63	Spain	1	1	0	0	1	0	0	0	0	0	0
	F	32	1.75	73	Belgium	4	3	0	1	2	0	0	4	0	0	1
a9	F	35	1.85	81	France	17	0	14	3	75	30	8	5	1	0	17
ɔ	F	27	1.83	76	Spain	21	13	6	4	30	13	6	8	1	0	10
·	F	22	1.75	73	Brazil	26	0	8	7	64	29	40	97	8	0	15
	F	22	1.75	63	Brazil	24	7	5	7	69	21	15	33	2	0	12
	F	29	1.8	76	Dominican Republic	6	6	0	0	3	0	3	0	2	0	0
ıez9	F	18	1.93	81	Uruguay	4	4	1	2	1	1	2	0	0	0	3

Figure 12: Adding new column of performance.

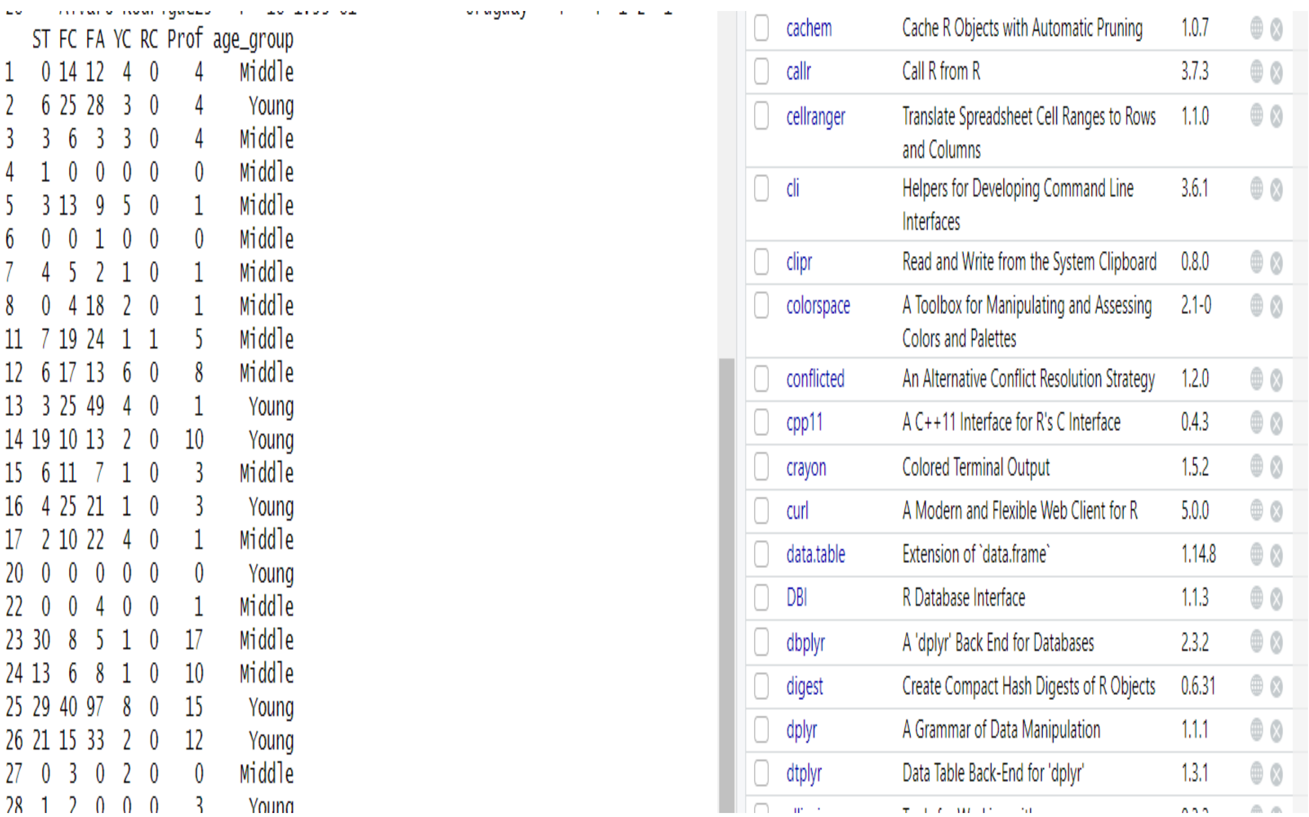
Then we try to categorize the age into a new variable to have a better understanding of the player's condition.

A new Column categorizing the age in which age is from 0 to equal than or less 25 are categorized as Young, age greater than 25 to equal or less than 40 are categorized as Middle, and age greater than 40 to equal or less than 100 are categorized as Old,

Code:

# Age grouping categorize

```
result$age_group <- cut(result$Age, breaks = c(0, 25, 40, 100), labels = c("Young", "Middle", "Old"))
result
```



```
   ST FC FA YC RC Prof age_group
1   0 14 12  4  0    4    Middle
2   6 25 28  3  0    4     Young
3   3  6  3  3  0    4    Middle
4   1  0  0  0  0    0    Middle
5   3 13  9  5  0    1    Middle
6   0  0  1  0  0    0    Middle
7   4  5  2  1  0    1    Middle
8   0  4 18  2  0    1    Middle
11  7 19 24  1  1    5    Middle
12  6 17 13  6  0    8    Middle
13  3 25 49  4  0    1     Young
14 19 10 13  2  0   10     Young
15  6 11  7  1  0    3    Middle
16  4 25 21  1  0    3     Young
17  2 10 22  4  0    1    Middle
20  0  0  0  0  0    0     Young
22  0  0  4  0  0    1    Middle
23 30  8  5  1  0   17    Middle
24 13  6  8  1  0   10    Middle
25 29 40 97  8  0   15     Young
26 21 15 33  2  0   12     Young
27  0  3  0  2  0    0    Middle
28  1  2  0  0  0    3     Young
```

□	cachem	Cache R Objects with Automatic Pruning	1.0.7
□	callr	Call R from R	3.7.3
□	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
□	cli	Helpers for Developing Command Line Interfaces	3.6.1
□	clipr	Read and Write from the System Clipboard	0.8.0
□	colorspace	A Toolbox for Manipulating and Assessing Colors and Palettes	2.1-0
□	conflicted	An Alternative Conflict Resolution Strategy	1.2.0
□	cpp11	A C++11 Interface for R's C Interface	0.4.3
□	crayon	Colored Terminal Output	1.5.2
□	curl	A Modern and Flexible Web Client for R	5.0.0
□	data.table	Extension of `data.frame`	1.14.8
□	DBI	R Database Interface	1.1.3
□	dbplyr	A 'dplyr' Back End for Databases	2.3.2
□	digest	Create Compact Hash Digests of R Objects	0.6.31
□	dplyr	A Grammar of Data Manipulation	1.1.1
□	dtplyr	Data Table Back-End for 'dplyr'	1.3.1

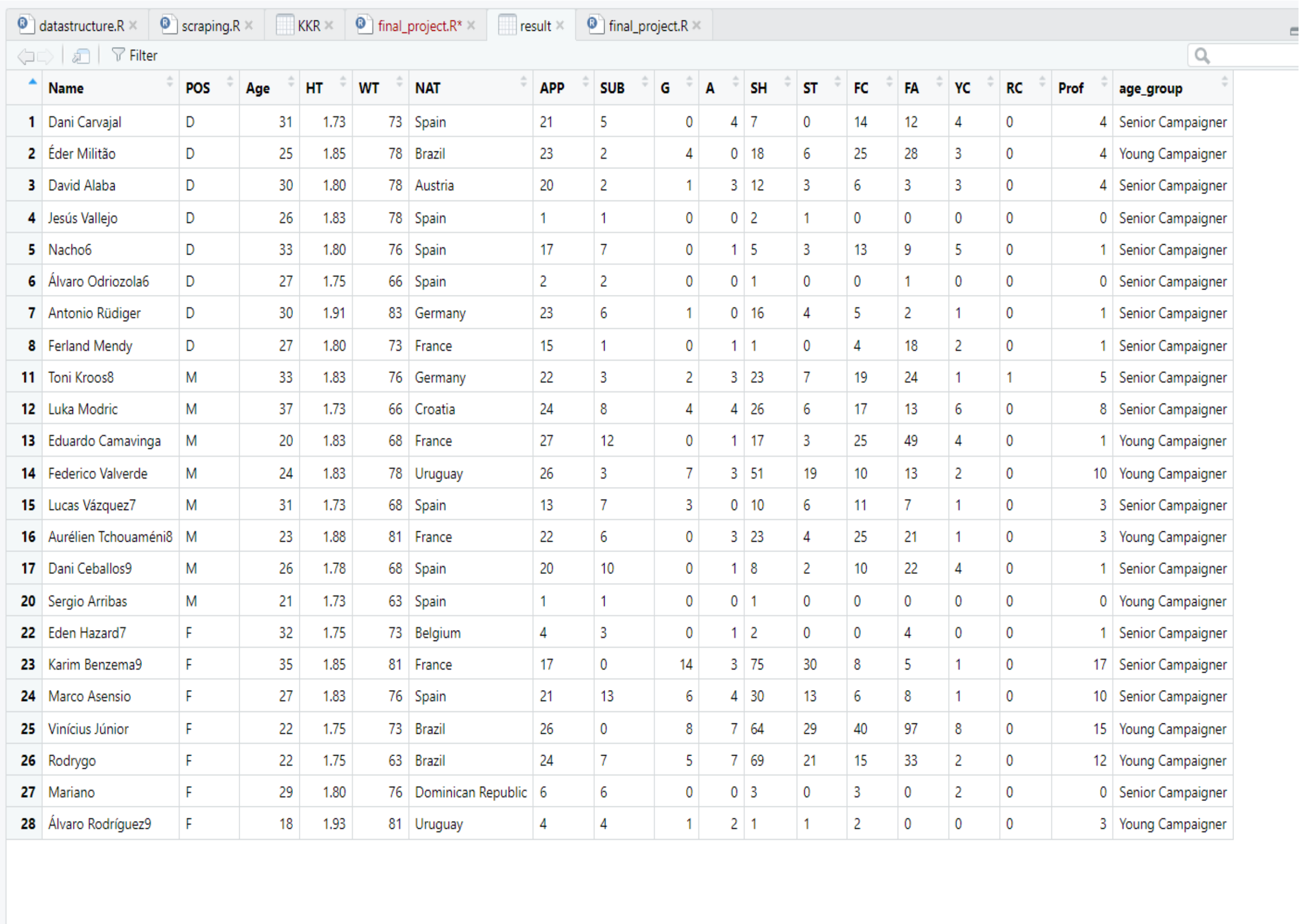
Figure 13: Age grouping categorize.

3. **Data Transformation:** In this phase, we need to transform some variables for better analysis of the dataset.

We need to transform the variables such as pos, HT, WT, NAT, AgeCat.

Code:

```
# Data transformation

result$POS <- factor(result$POS, ordered = TRUE)

result$HT <- as.numeric(result$HT)
result$WT <- as.numeric(result$WT)

result$NAT <- factor(result$NAT, ordered = TRUE)

result$age_group <- factor(result$age_group,
                    levels =c("Young", "Middle", "old"),labels=c("Young Campaigner","Senior
Campaigner","Old Campaigner"))

result
```



	Name	POS	Age	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC	Prof	age_group
1	Dani Carvajal	D	31	1.73	73	Spain	21	5	0	4	7	0	14	12	4	0	4	Senior Campaigner
2	Éder Militão	D	25	1.85	78	Brazil	23	2	4	0	18	6	25	28	3	0	4	Young Campaigner
3	David Alaba	D	30	1.80	78	Austria	20	2	1	3	12	3	6	3	3	0	4	Senior Campaigner
4	Jesús Vallejo	D	26	1.83	78	Spain	1	1	0	0	2	1	0	0	0	0	0	Senior Campaigner
5	Nacho6	D	33	1.80	76	Spain	17	7	0	1	5	3	13	9	5	0	1	Senior Campaigner
6	Álvaro Odriozola6	D	27	1.75	66	Spain	2	2	0	0	1	0	0	1	0	0	0	Senior Campaigner
7	Antonio Rüdiger	D	30	1.91	83	Germany	23	6	1	0	16	4	5	2	1	0	1	Senior Campaigner
8	Ferland Mendy	D	27	1.80	73	France	15	1	0	1	1	0	4	18	2	0	1	Senior Campaigner
11	Toni Kroos8	M	33	1.83	76	Germany	22	3	2	3	23	7	19	24	1	1	5	Senior Campaigner
12	Luka Modric	M	37	1.73	66	Croatia	24	8	4	4	26	6	17	13	6	0	8	Senior Campaigner
13	Eduardo Camavinga	M	20	1.83	68	France	27	12	0	1	17	3	25	49	4	0	1	Young Campaigner
14	Federico Valverde	M	24	1.83	78	Uruguay	26	3	7	3	51	19	10	13	2	0	10	Young Campaigner
15	Lucas Vázquez7	M	31	1.73	68	Spain	13	7	3	0	10	6	11	7	1	0	3	Senior Campaigner
16	Aurélien Tchouaméni8	M	23	1.88	81	France	22	6	0	3	23	4	25	21	1	0	3	Young Campaigner
17	Dani Ceballos9	M	26	1.78	68	Spain	20	10	0	1	8	2	10	22	4	0	1	Senior Campaigner
20	Sergio Arribas	M	21	1.73	63	Spain	1	1	0	0	1	0	0	0	0	0	0	Young Campaigner
22	Eden Hazard7	F	32	1.75	73	Belgium	4	3	0	1	2	0	0	4	0	0	1	Senior Campaigner
23	Karim Benzema9	F	35	1.85	81	France	17	0	14	3	75	30	8	5	1	0	17	Senior Campaigner
24	Marco Asensio	F	27	1.83	76	Spain	21	13	6	4	30	13	6	8	1	0	10	Senior Campaigner
25	Vinícius Júnior	F	22	1.75	73	Brazil	26	0	8	7	64	29	40	97	8	0	15	Young Campaigner
26	Rodrygo	F	22	1.75	63	Brazil	24	7	5	7	69	21	15	33	2	0	12	Young Campaigner
27	Mariano	F	29	1.80	76	Dominican Republic	6	6	0	0	3	0	3	0	2	0	0	Senior Campaigner
28	Álvaro Rodríguez9	F	18	1.93	81	Uruguay	4	4	1	2	1	1	2	0	0	0	3	Young Campaigner

Figure 14: Data transformation.

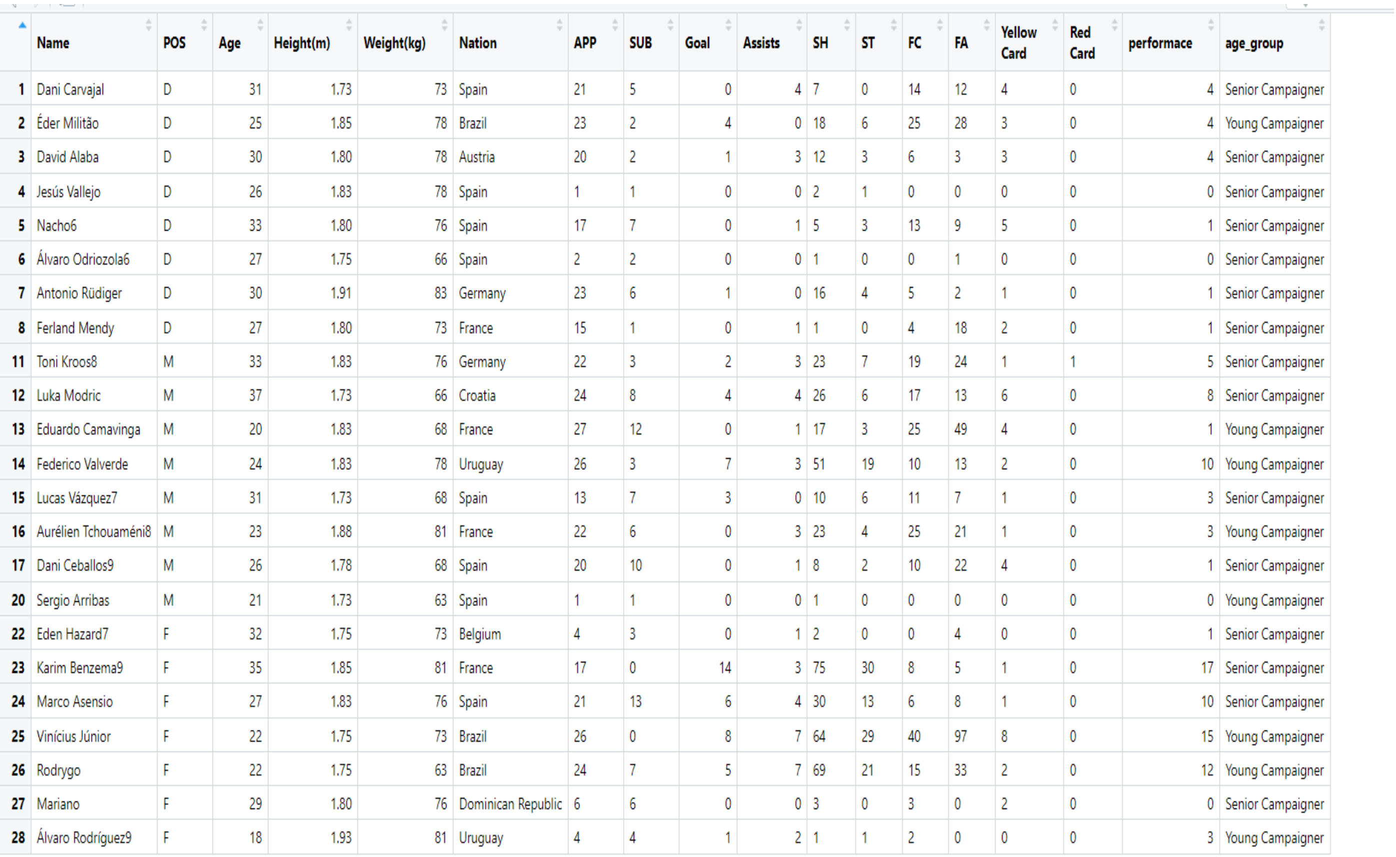
Some of the column names are pretty hard to understand, for this reason, we need to change some of the column names for understanding the database more thoroughly.

Changing some of the column names,

Code:

#Rename some columns

```
colnames(result)[5] <- "Weight(kg)"
colnames(result)[9] <- "Goal"
colnames(result)[10] <- "Assists"
colnames(result)[6] <- "Nation"
colnames(result)[15] <- "Yellow Card"
colnames(result)[16] <- "Red Card"
colnames(result)[17] <- "performance"
```



	Name	POS	Age	Height(m)	Weight(kg)	Nation	APP	SUB	Goal	Assists	SH	ST	FC	FA	Yellow Card	Red Card	performace	age_group
1	Dani Carvajal	D	31	1.73	73	Spain	21	5	0	4	7	0	14	12	4	0	4	Senior Campaigner
2	Éder Militão	D	25	1.85	78	Brazil	23	2	4	0	18	6	25	28	3	0	4	Young Campaigner
3	David Alaba	D	30	1.80	78	Austria	20	2	1	3	12	3	6	3	3	0	4	Senior Campaigner
4	Jesús Vallejo	D	26	1.83	78	Spain	1	1	0	0	2	1	0	0	0	0	0	Senior Campaigner
5	Nacho6	D	33	1.80	76	Spain	17	7	0	1	5	3	13	9	5	0	1	Senior Campaigner
6	Álvaro Odriozola6	D	27	1.75	66	Spain	2	2	0	0	1	0	0	1	0	0	0	Senior Campaigner
7	Antonio Rüdiger	D	30	1.91	83	Germany	23	6	1	0	16	4	5	2	1	0	1	Senior Campaigner
8	Ferland Mendy	D	27	1.80	73	France	15	1	0	1	1	0	4	18	2	0	1	Senior Campaigner
11	Toni Kroos8	M	33	1.83	76	Germany	22	3	2	3	23	7	19	24	1	1	5	Senior Campaigner
12	Luka Modric	M	37	1.73	66	Croatia	24	8	4	4	26	6	17	13	6	0	8	Senior Campaigner
13	Eduardo Camavinga	M	20	1.83	68	France	27	12	0	1	17	3	25	49	4	0	1	Young Campaigner
14	Federico Valverde	M	24	1.83	78	Uruguay	26	3	7	3	51	19	10	13	2	0	10	Young Campaigner
15	Lucas Vázquez7	M	31	1.73	68	Spain	13	7	3	0	10	6	11	7	1	0	3	Senior Campaigner
16	Aurélien Tchouaméni8	M	23	1.88	81	France	22	6	0	3	23	4	25	21	1	0	3	Young Campaigner
17	Dani Ceballos9	M	26	1.78	68	Spain	20	10	0	1	8	2	10	22	4	0	1	Senior Campaigner
20	Sergio Arribas	M	21	1.73	63	Spain	1	1	0	0	1	0	0	0	0	0	0	Young Campaigner
22	Eden Hazard7	F	32	1.75	73	Belgium	4	3	0	1	2	0	0	4	0	0	1	Senior Campaigner
23	Karim Benzema9	F	35	1.85	81	France	17	0	14	3	75	30	8	5	1	0	17	Senior Campaigner
24	Marco Asensio	F	27	1.83	76	Spain	21	13	6	4	30	13	6	8	1	0	10	Senior Campaigner
25	Vinícius Júnior	F	22	1.75	73	Brazil	26	0	8	7	64	29	40	97	8	0	15	Young Campaigner
26	Rodrygo	F	22	1.75	63	Brazil	24	7	5	7	69	21	15	33	2	0	12	Young Campaigner
27	Mariano	F	29	1.80	76	Dominican Republic	6	6	0	0	3	0	3	0	2	0	0	Senior Campaigner
28	Álvaro Rodríguez9	F	18	1.93	81	Uruguay	4	4	1	2	1	1	2	0	0	0	3	Young Campaigner

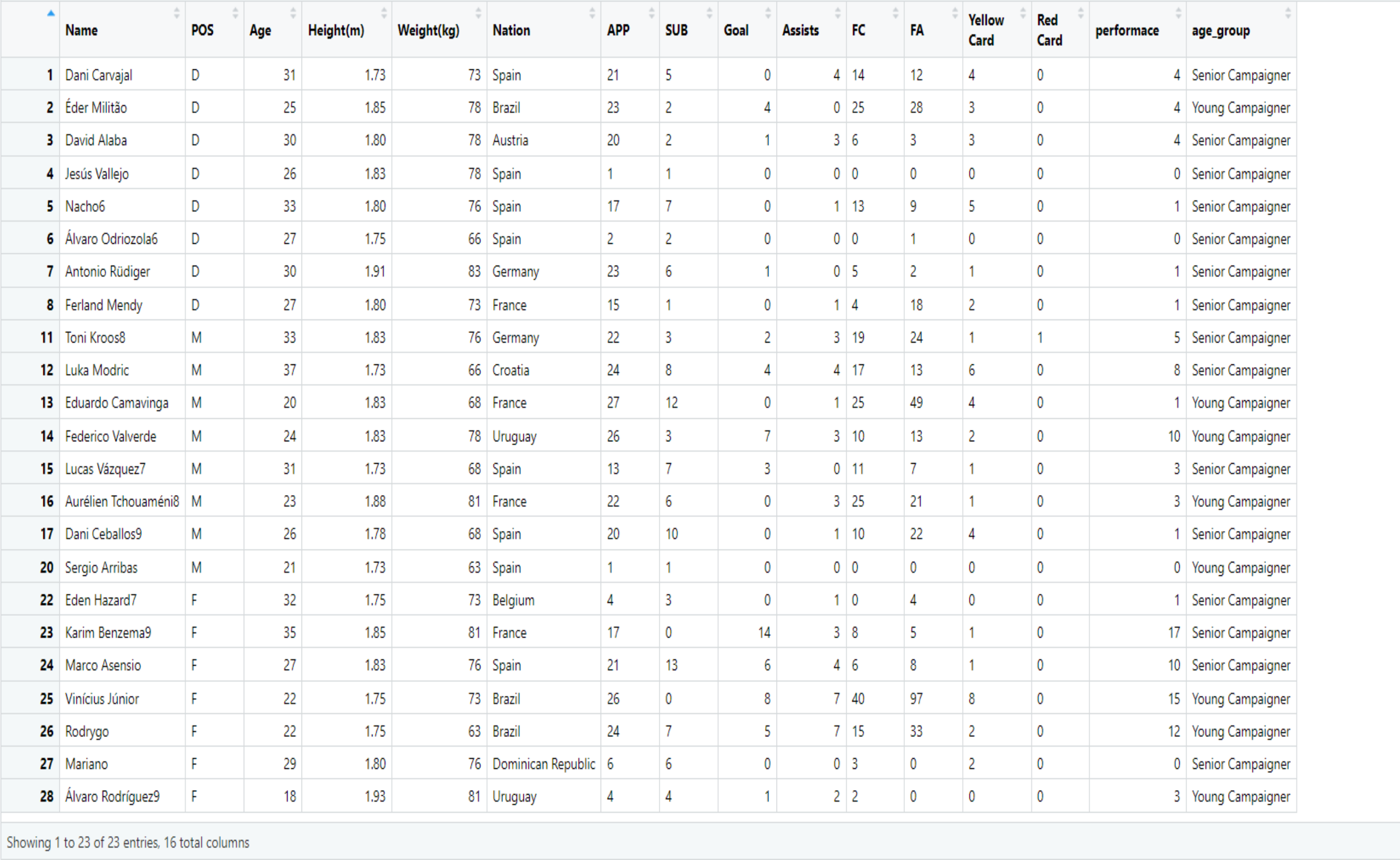
Figure 15: Rename some columns.

4. **Data Reduction:** In our dataset, we can see that some columns are not necessary for analysis. So, we remove those columns from the dataset.

Code:

#Data reduction

```
result <- subset(result, select = -c(ST))

result <- subset(result, select = -c(SH))

View(result)
```

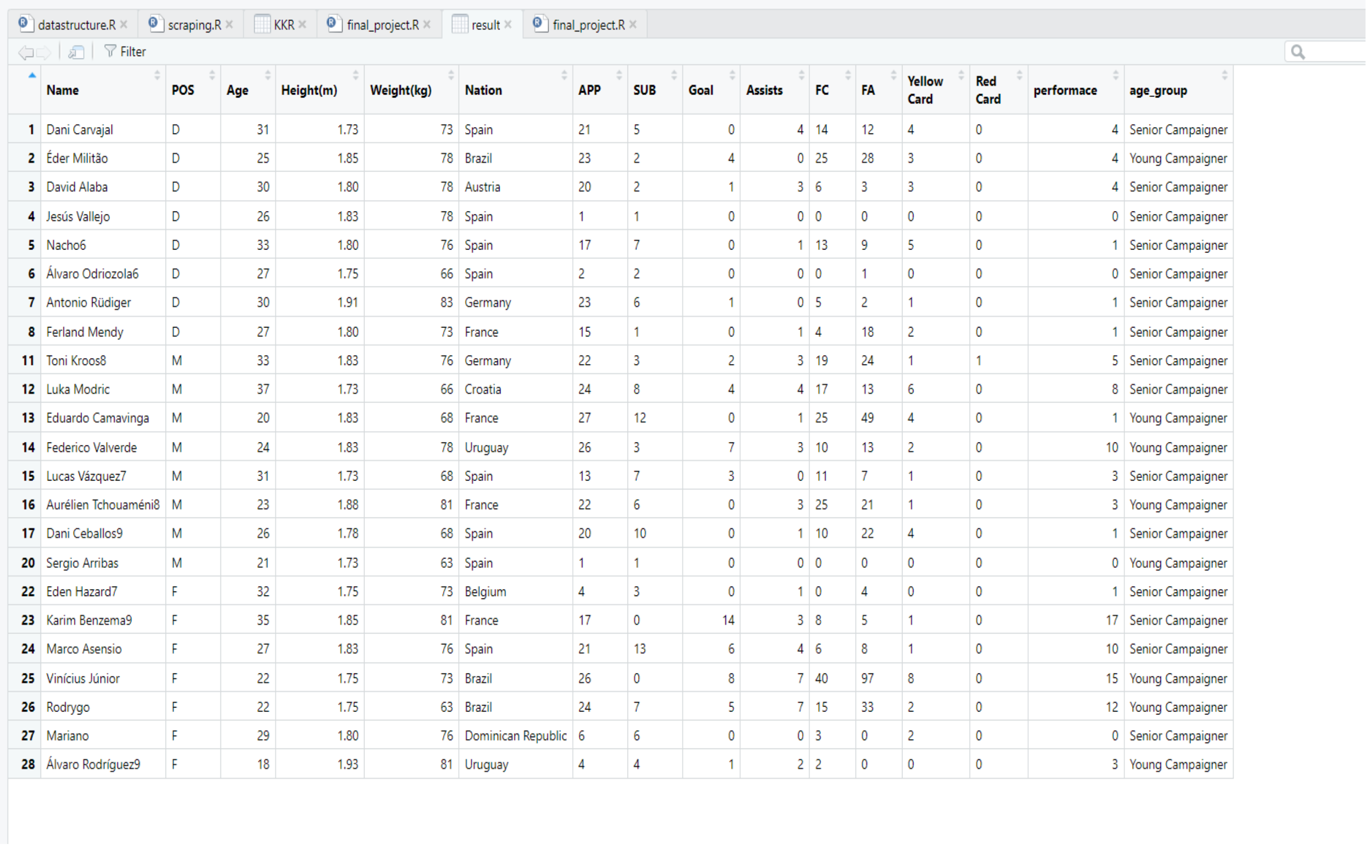


	Name	POS	Age	Height(m)	Weight(kg)	Nation	APP	SUB	Goal	Assists	FC	FA	Yellow Card	Red Card	performace	age_group
1	Dani Carvajal	D	31	1.73	73	Spain	21	5	0	4	14	12	4	0	4	Senior Campaigner
2	Éder Militão	D	25	1.85	78	Brazil	23	2	4	0	25	28	3	0	4	Young Campaigner
3	David Alaba	D	30	1.80	78	Austria	20	2	1	3	6	3	3	0	4	Senior Campaigner
4	Jesús Vallejo	D	26	1.83	78	Spain	1	1	0	0	0	0	0	0	0	Senior Campaigner
5	Nacho6	D	33	1.80	76	Spain	17	7	0	1	13	9	5	0	1	Senior Campaigner
6	Álvaro Odriozola6	D	27	1.75	66	Spain	2	2	0	0	0	1	0	0	0	Senior Campaigner
7	Antonio Rüdiger	D	30	1.91	83	Germany	23	6	1	0	5	2	1	0	1	Senior Campaigner
8	Ferland Mendy	D	27	1.80	73	France	15	1	0	1	4	18	2	0	1	Senior Campaigner
11	Toni Kroos8	M	33	1.83	76	Germany	22	3	2	3	19	24	1	1	5	Senior Campaigner
12	Luka Modric	M	37	1.73	66	Croatia	24	8	4	4	17	13	6	0	8	Senior Campaigner
13	Eduardo Camavinga	M	20	1.83	68	France	27	12	0	1	25	49	4	0	1	Young Campaigner
14	Federico Valverde	M	24	1.83	78	Uruguay	26	3	7	3	10	13	2	0	10	Young Campaigner
15	Lucas Vázquez7	M	31	1.73	68	Spain	13	7	3	0	11	7	1	0	3	Senior Campaigner
16	Aurélien Tchouaméni8	M	23	1.88	81	France	22	6	0	3	25	21	1	0	3	Young Campaigner
17	Dani Ceballos9	M	26	1.78	68	Spain	20	10	0	1	10	22	4	0	1	Senior Campaigner
20	Sergio Arribas	M	21	1.73	63	Spain	1	1	0	0	0	0	0	0	0	Young Campaigner
22	Eden Hazard7	F	32	1.75	73	Belgium	4	3	0	1	0	4	0	0	1	Senior Campaigner
23	Karim Benzema9	F	35	1.85	81	France	17	0	14	3	8	5	1	0	17	Senior Campaigner
24	Marco Asensio	F	27	1.83	76	Spain	21	13	6	4	6	8	1	0	10	Senior Campaigner
25	Vinícius Júnior	F	22	1.75	73	Brazil	26	0	8	7	40	97	8	0	15	Young Campaigner
26	Rodrygo	F	22	1.75	63	Brazil	24	7	5	7	15	33	2	0	12	Young Campaigner
27	Mariano	F	29	1.80	76	Dominican Republic	6	6	0	0	3	0	2	0	0	Senior Campaigner
28	Álvaro Rodríguez9	F	18	1.93	81	Uruguay	4	4	1	2	2	0	0	0	3	Young Campaigner

Showing 1 to 23 of 23 entries, 16 total columns

Figure 16: Data reduction.

5. **Data Discretization:** No discretization is needed for this dataset as it is already in a better shape. So, we skip this process and move on to descriptive statistics.



	Name	POS	Age	Height(m)	Weight(kg)	Nation	APP	SUB	Goal	Assists	FC	FA	Yellow Card	Red Card	performace	age_group
1	Dani Carvajal	D	31	1.73	73	Spain	21	5	0	4	14	12	4	0	4	Senior Campaigner
2	Éder Militão	D	25	1.85	78	Brazil	23	2	4	0	25	28	3	0	4	Young Campaigner
3	David Alaba	D	30	1.80	78	Austria	20	2	1	3	6	3	3	0	4	Senior Campaigner
4	Jesús Vallejo	D	26	1.83	78	Spain	1	1	0	0	0	0	0	0	0	Senior Campaigner
5	Nacho6	D	33	1.80	76	Spain	17	7	0	1	13	9	5	0	1	Senior Campaigner
6	Álvaro Odriozola6	D	27	1.75	66	Spain	2	2	0	0	0	1	0	0	0	Senior Campaigner
7	Antonio Rüdiger	D	30	1.91	83	Germany	23	6	1	0	5	2	1	0	1	Senior Campaigner
8	Ferland Mendy	D	27	1.80	73	France	15	1	0	1	4	18	2	0	1	Senior Campaigner
11	Toni Kroos8	M	33	1.83	76	Germany	22	3	2	3	19	24	1	1	5	Senior Campaigner
12	Luka Modric	M	37	1.73	66	Croatia	24	8	4	4	17	13	6	0	8	Senior Campaigner
13	Eduardo Camavinga	M	20	1.83	68	France	27	12	0	1	25	49	4	0	1	Young Campaigner
14	Federico Valverde	M	24	1.83	78	Uruguay	26	3	7	3	10	13	2	0	10	Young Campaigner
15	Lucas Vázquez7	M	31	1.73	68	Spain	13	7	3	0	11	7	1	0	3	Senior Campaigner
16	Aurélien Tchouaméni8	M	23	1.88	81	France	22	6	0	3	25	21	1	0	3	Young Campaigner
17	Dani Ceballos9	M	26	1.78	68	Spain	20	10	0	1	10	22	4	0	1	Senior Campaigner
20	Sergio Arribas	M	21	1.73	63	Spain	1	1	0	0	0	0	0	0	0	Young Campaigner
22	Eden Hazard7	F	32	1.75	73	Belgium	4	3	0	1	0	4	0	0	1	Senior Campaigner
23	Karim Benzema9	F	35	1.85	81	France	17	0	14	3	8	5	1	0	17	Senior Campaigner
24	Marco Asensio	F	27	1.83	76	Spain	21	13	6	4	6	8	1	0	10	Senior Campaigner
25	Vinícius Júnior	F	22	1.75	73	Brazil	26	0	8	7	40	97	8	0	15	Young Campaigner
26	Rodrygo	F	22	1.75	63	Brazil	24	7	5	7	15	33	2	0	12	Young Campaigner
27	Mariano	F	29	1.80	76	Dominican Republic	6	6	0	0	3	0	2	0	0	Senior Campaigner
28	Álvaro Rodríguez9	F	18	1.93	81	Uruguay	4	4	1	2	2	0	0	0	3	Young Campaigner

Figure 17: The total data frame after applying Pre-processing.

Descriptive Statistics:

Now, we are going to compute various descriptive statistics parameters for our dataset.

Firstly, let’s try to inspect the central tendency for the various variables of our dataset.

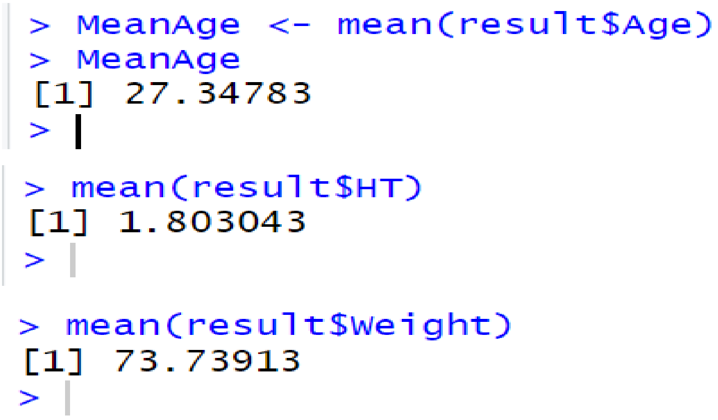
- **MEAN:** The mean is a measure of central tendency that is calculated by adding together all of the observations and dividing by the number of observations.

  Mean value of all player’s ages, weights and heights.

  Code:

  ```
  MeanAge <- mean(result$Age)

  MeanAge


  # Check the data type of my_object

  class(result$HT)
  ```

```
# If the data type is not numeric or logical, convert it to a numeric or logical object
result$HT <- as.numeric(result$HT)

# Check for missing values in my_object
is.na(result$HT)

# If there are missing values, remove them using na.omit()
result$HT <- na.omit(result$HT)

# Calculate the mean of my_object
mean(result$HT)

# Check the data type of my_object
class(result$Weight)

# If the data type is not numeric or logical, convert it to a numeric or logical object
result$WT <- as.numeric(result$Weight)

# Check for missing values in my_object
is.na(result$Weight)

# If there are missing values, remove them using na.omit()
result$WT <- na.omit(result$Weight)

# Calculate the mean of my_object
mean(result$Weight)
```



```
> MeanAge <- mean(result$Age)
> MeanAge
[1] 27.34783
> 
> mean(result$HT)
[1] 1.803043
> 
> mean(result$Weight)
[1] 73.73913
> 
```

Figure 18: Mean value of all player's ages, weights and heights.

- **<u>MEDIAN:</u>** The median is another measure of central tendency but one that cannot be directly calculated. Instead, you make a sorted list of all of the observations in the sample and then go halfway up that list.

Now we calculate the median for the number of fouls committed and fouls suffered,

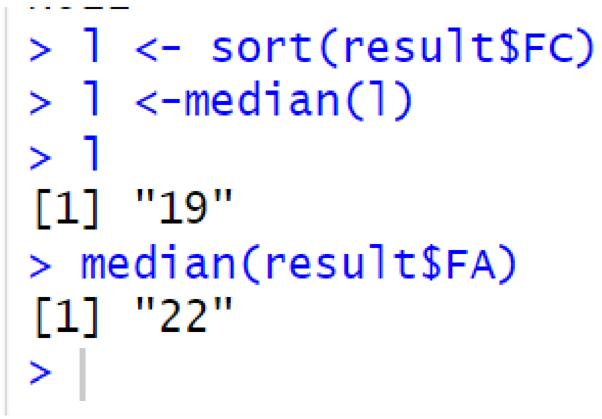
<u>Code:</u>

l <- sort(result$FC)

l <-median(l)

l

median(result$FA)



```
> l <- sort(result$FC)
> l <-median(l)
> l
[1] "19"
> median(result$FA)
[1] "22"
> 
```

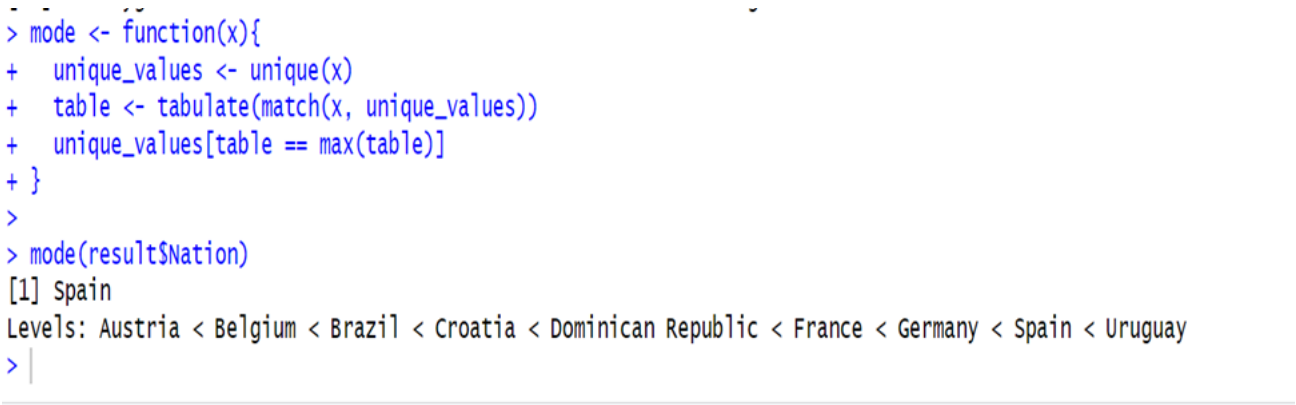
Figure 19: Median value for the number of fouls committed and fouls suffered.

- **MODE:** The mode is another measure of central tendency. The mode is the value that occurs most often in a sample of data.

As the mode doesn’t have a built-in function, we first implement the function.

Code:

```
mode <- function(x){
  unique_values <- unique(x)
  table <- tabulate(match(x, unique_values))
  unique_values[table == max(table)]
}
mode(result$Nation)
```



```
> mode <- function(x){
+   unique_values <- unique(x)
+   table <- tabulate(match(x, unique_values))
+   unique_values[table == max(table)]
+ }
>
> mode(result$Nation)
[1] Spain
Levels: Austria < Belgium < Brazil < Croatia < Dominican Republic < France < Germany < Spain < Uruguay
>
```

Figure 20: Mode value for the players Nation.

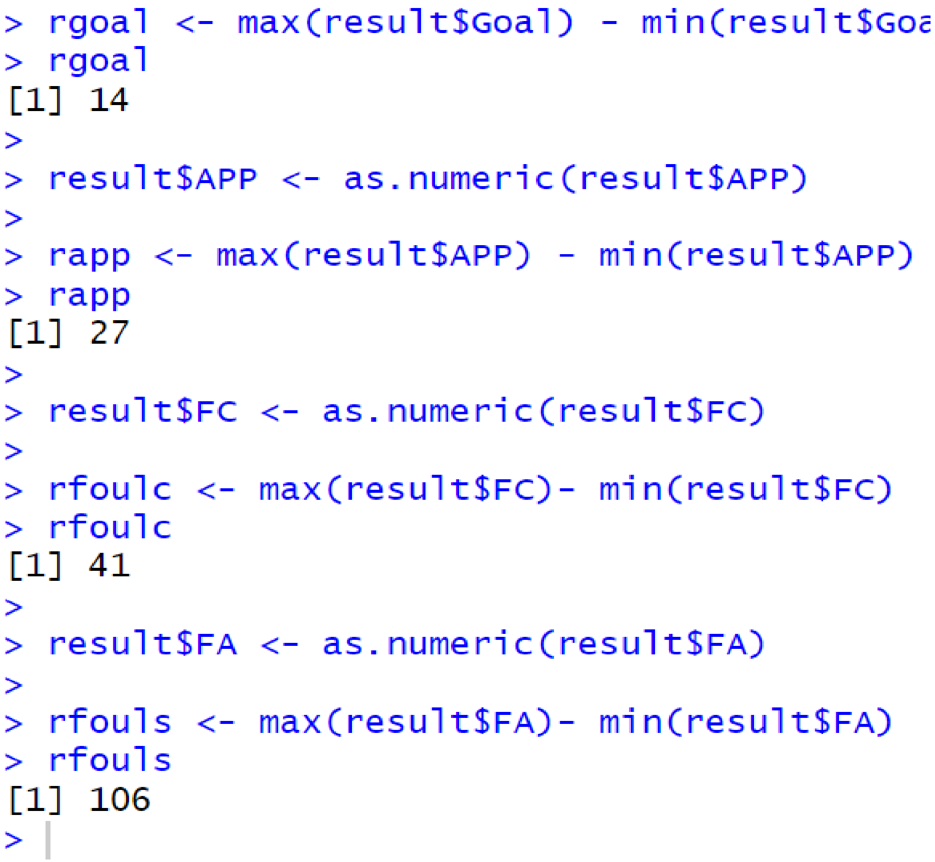
- **Range:** The range is a measure of dispersion—how spread out a bunch of numbers in a sample are—calculated by subtracting the lowest value from the highest value.

Now we calculate the range of variables.

Code:

```
rgoal <- max(result$Goal) - min(result$Goal)
rgoal
result$APP <- as.numeric(result$APP)
```

```
rapp <- max(result$APP) - min(result$APP)

rapp

result$FC <- as.numeric(result$FC)


rfoulc <- max(result$FC)- min(result$FC)

rfoulc

result$FA <- as.numeric(result$FA)


rfouls <- max(result$FA)- min(result$FA)

rfouls
```



```
> rgoal <- max(result$Goal) - min(result$Goa
> rgoal
[1] 14
>
> result$APP <- as.numeric(result$APP)
>
> rapp <- max(result$APP) - min(result$APP)
> rapp
[1] 27
>
> result$FC <- as.numeric(result$FC)
>
> rfoulc <- max(result$FC)- min(result$FC)
> rfoulc
[1] 41
>
> result$FA <- as.numeric(result$FA)
>
> rfouls <- max(result$FA)- min(result$FA)
> rfouls
[1] 106
> 
```

Figure 21: Range values of variables.

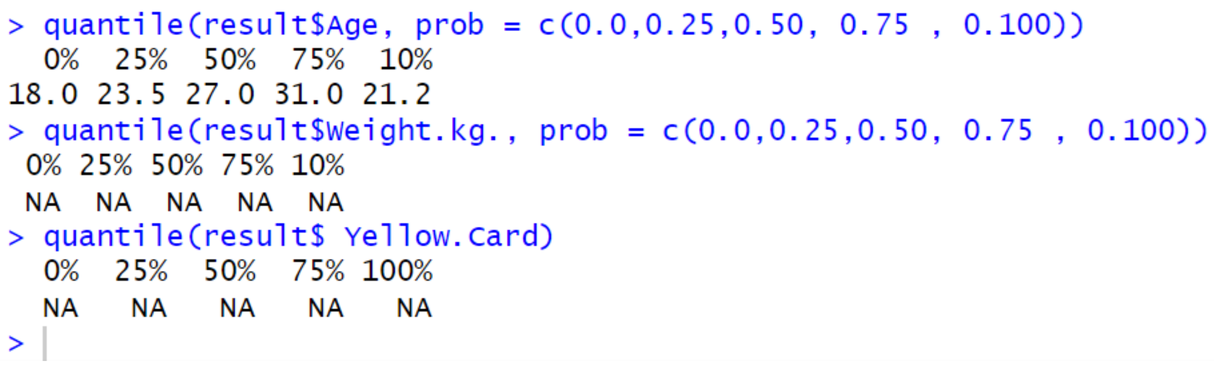
- **Quartile & Percentile:**
- **Quartiles** are values that separate the data into four equal parts. The quartiles (Q_0, Q_1, Q_2, Q_3, Q4) are the values that separate each quarter.
- **Percentiles** are values that separate the data into 100 equal parts.

Code:

quantile(result$Age, prob = c(0.0,0.25,0.50, 0.75 , 0.100))

quantile(result$Weight.kg., prob = c(0.0,0.25,0.50, 0.75 , 0.100))

quantile(result$ Yellow.Card)



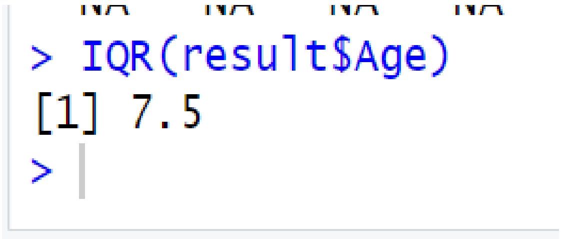
```
> quantile(result$Age, prob = c(0.0,0.25,0.50, 0.75 , 0.100))
  0%  25%  50%  75%  10% 
18.0 23.5 27.0 31.0 21.2 
> quantile(result$Weight.kg., prob = c(0.0,0.25,0.50, 0.75 , 0.100))
 0% 25% 50% 75% 10% 
 NA  NA  NA  NA  NA 
> quantile(result$ Yellow.Card)
  0%  25%  50%  75% 100% 
  NA   NA   NA   NA   NA 
> 
```

Figure 22: Quartile & Percentile values of variables.

- **Interquartile Range:** Interquartile range is the difference between the first and third quartiles (Q1 and Q3).

Code:

IQR(result$Age)



```
> IQR(result$Age)
[1] 7.5
> 
```

Figure 23: Interquartile Range values of player's age.

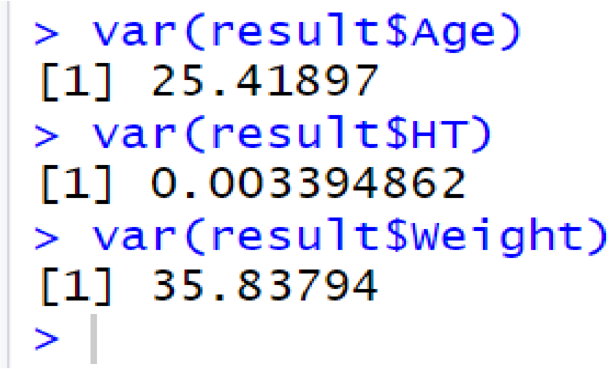
- **Variance:** The variance is a measure of dispersion.

Code:

var(result$Age)

var(result$HT)

var(result$Weight)



```
> var(result$Age)
[1] 25.41897
> var(result$HT)
[1] 0.003394862
> var(result$Weight)
[1] 35.83794
> 
```

Figure 24: Variance value of all player's ages, weights and heights.

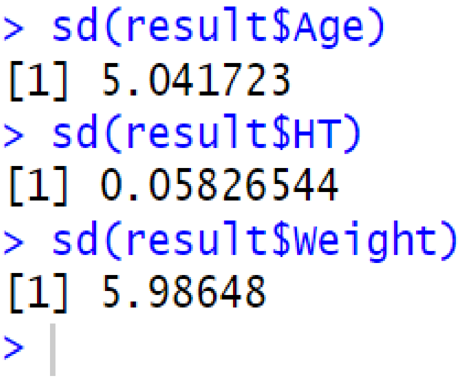
- **Standard Deviation:** The standard deviation is simply the square root of the variance. Standard deviation measures how far a 'typical' observation is from the average of the data.

Code:

sd(result$Age)

sd(result$HT)

sd(result$Weight)



```
> sd(result$Age)
[1] 5.041723
> sd(result$HT)
[1] 0.05826544
> sd(result$Weight)
[1] 5.98648
> 
```

Figure 25: Standard Deviation value of all player's ages, weights and heights.

- **Normal Distribution:** In a normal distribution, the values are concentrated around a given value (i.e., the *mean* value, and the value of the standard deviation from the mean).

Code:

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

hist(x)

z = rnorm(result$Goal, mean = mean(result$Goal),sd = sd(result$Goal) )

hist(z)

y = dnorm(result$APP , mean = mean(result$APP), sd= sd(result$APP))

plot(result$APP,y)
```

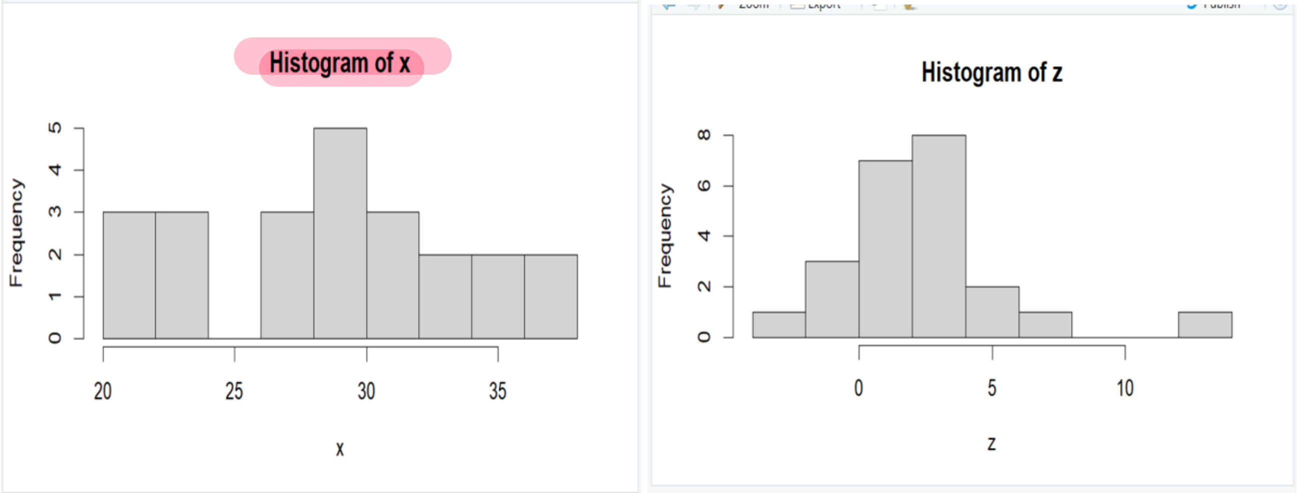


Figure 26: Normal Distribution values of variable x and z.

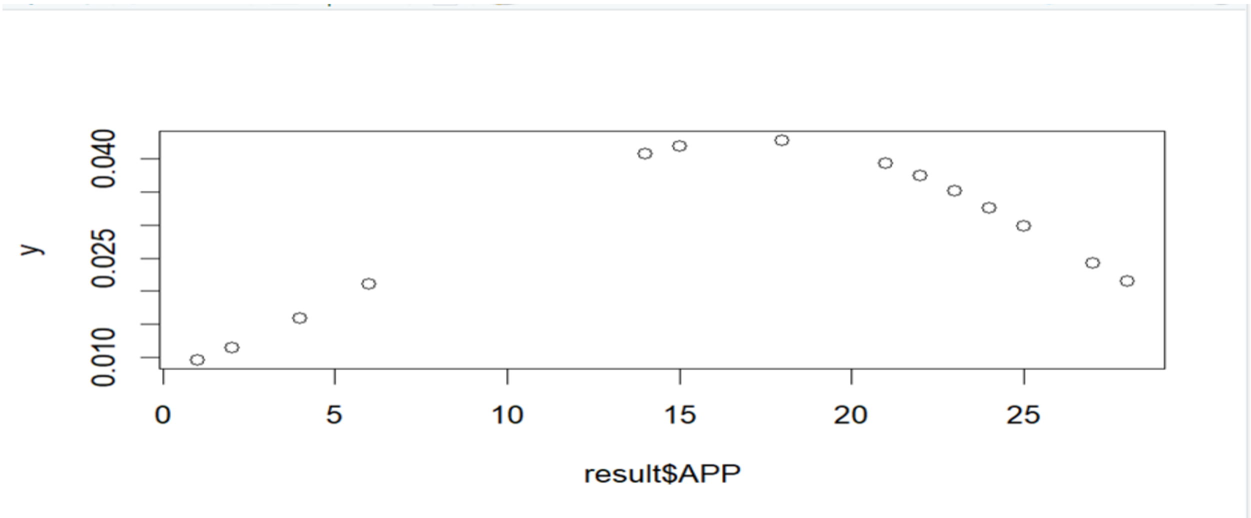


Figure 27: Normal Distribution value of y.

Data Visualization: It involves presenting data in a graphical or visual format that is easily understandable to the viewer. Data visualization is used to summarize, explore, and communicate data effectively. It can be used to identify patterns, trends, and relationships in the data that might not be immediately apparent from looking at the raw data. Data visualization can be performed using various tools and techniques, including graphs, charts, maps etc.

Codes:

```
library(ggplot2)
```

#1) First let's draw a scatter plot of Appearance vs Goal for each team,

Code:

```
ggplot(result, 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"))+
facet_wrap(~age_group)
```

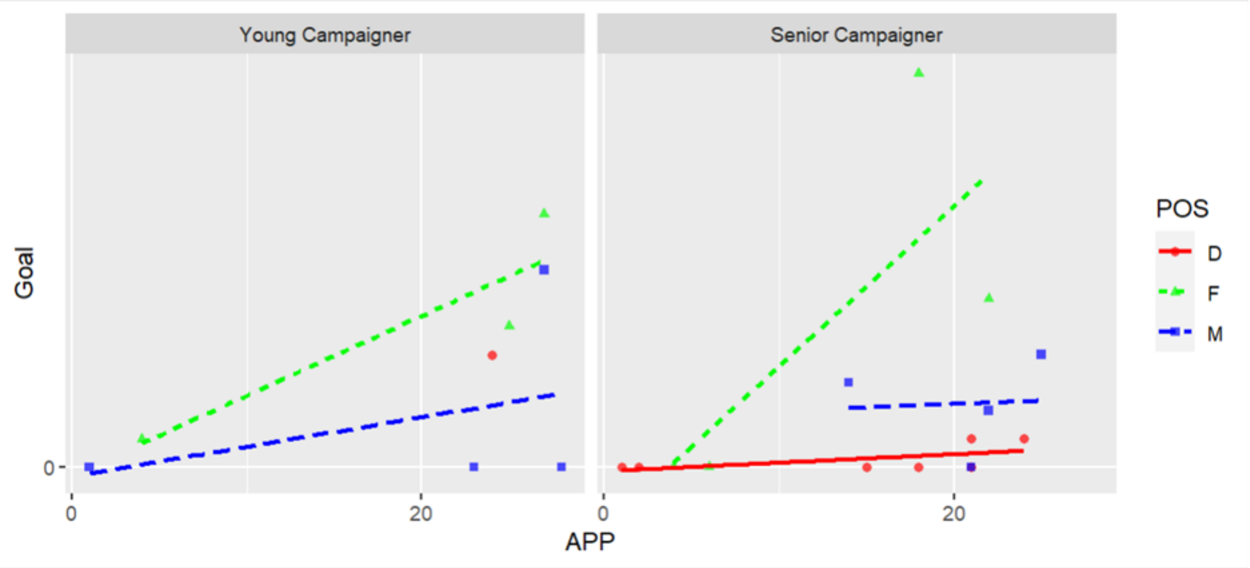


Figure 28: From this scatter plot, we can understand that the player with more appearances started to score more goals. In the Young Campaigner side, the forward with more appearances started to deliver more goals, and in the Senior Campaigner side, forwards started to show extra ordinary numbers with more appearances.

#2) Now we see a scatter plot for Defenders Appearance vs Fouls Committed,

Code:

```
ggplot(result, aes(x = APP, y= FC, 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"))+
  facet_wrap(~age_group)
```

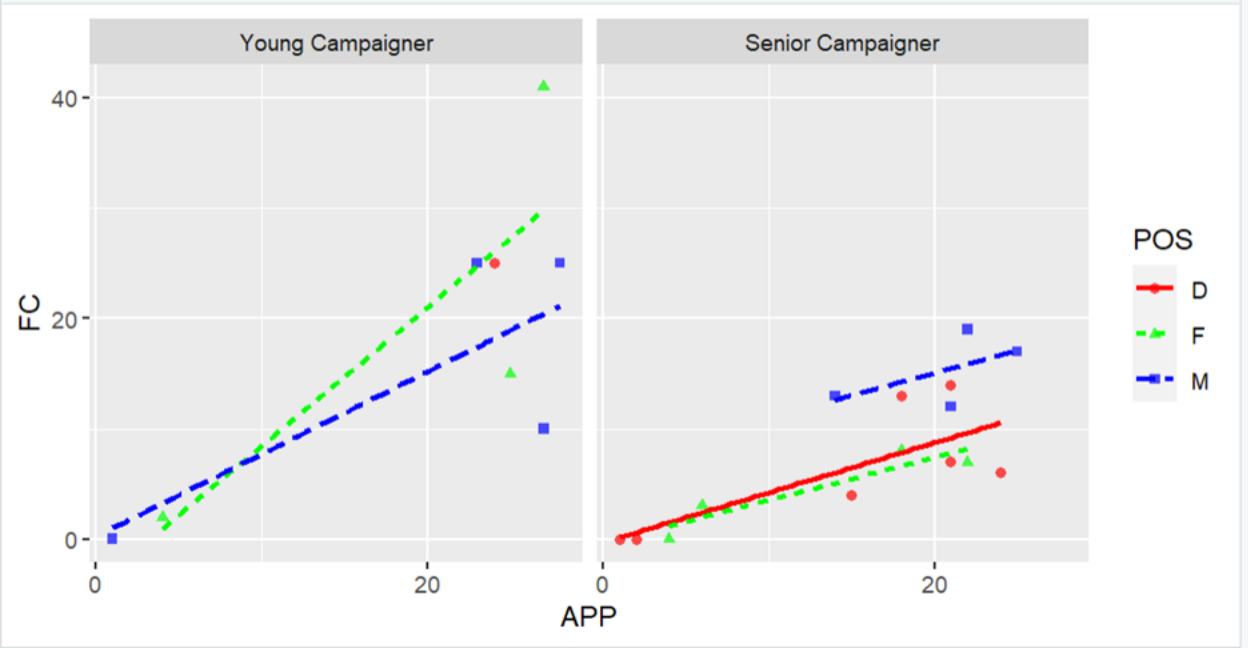


Figure 29: In this plot, we can see that with more appearances Young Campaigner defenders started to be more aggressive than Senior Campaigner Defenders. But most of the attacks of the Young Campaigner side come from the Midfielders.

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

Code:

```
library(ggpie)
library(dplyr)

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

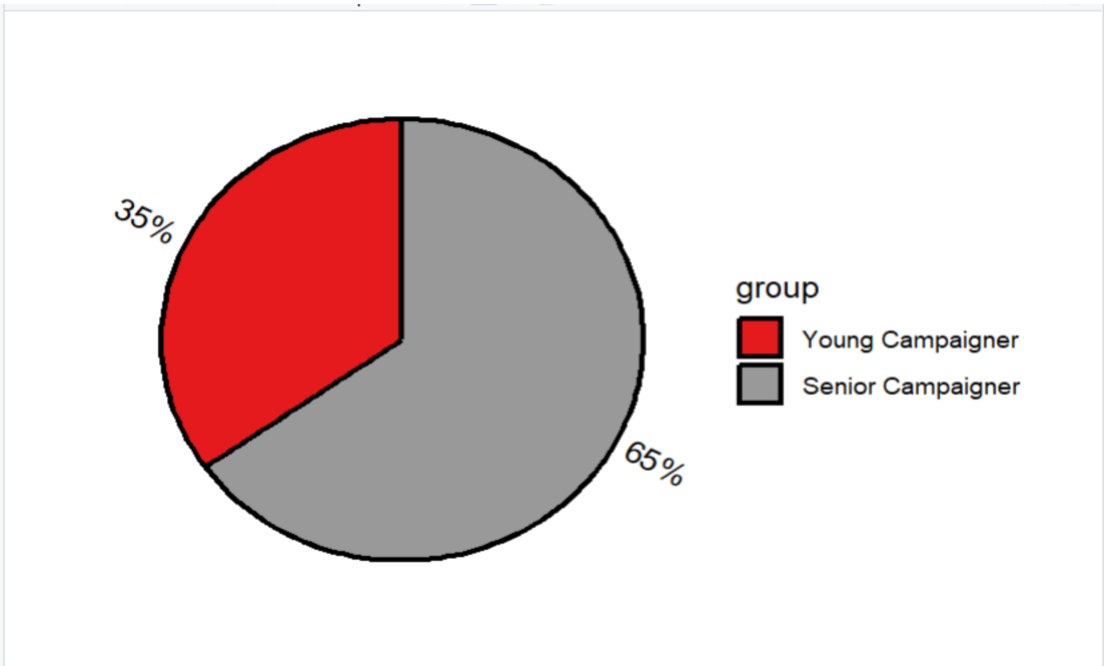


Figure 30: In this pie chart we can see that the majority of the players belong to the senior campaigner category which means the team is filled with experienced players.

#4) Furthermore, we try to identify the greatest number of players from and individual country.

Code:

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

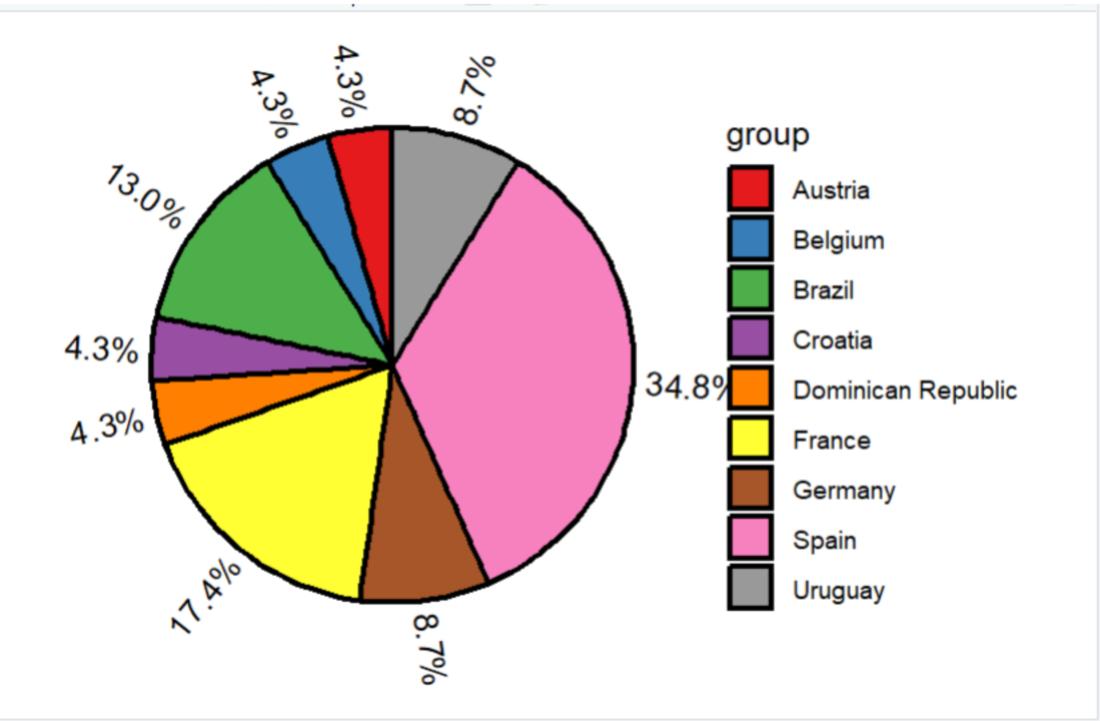


Figure 31: From the pie chart we can identify that, most of the players are from Spain which is 34.8%. The next most population is from France. We also get some ideas of their playing style, as most of the players are from Spain, they prefer tiki taka.

#5) Now the most important part of the visualization. We need to see the contribution of the forwards for the respective team,

Code:

```
library(ggplot2)
ggplot(result,aes(x=performace, fill=age_group))+
 geom_bar()+
 labs(title = "Contribution Of Forwards", x ="performance", y="Frequency")+
 coord_flip()
```

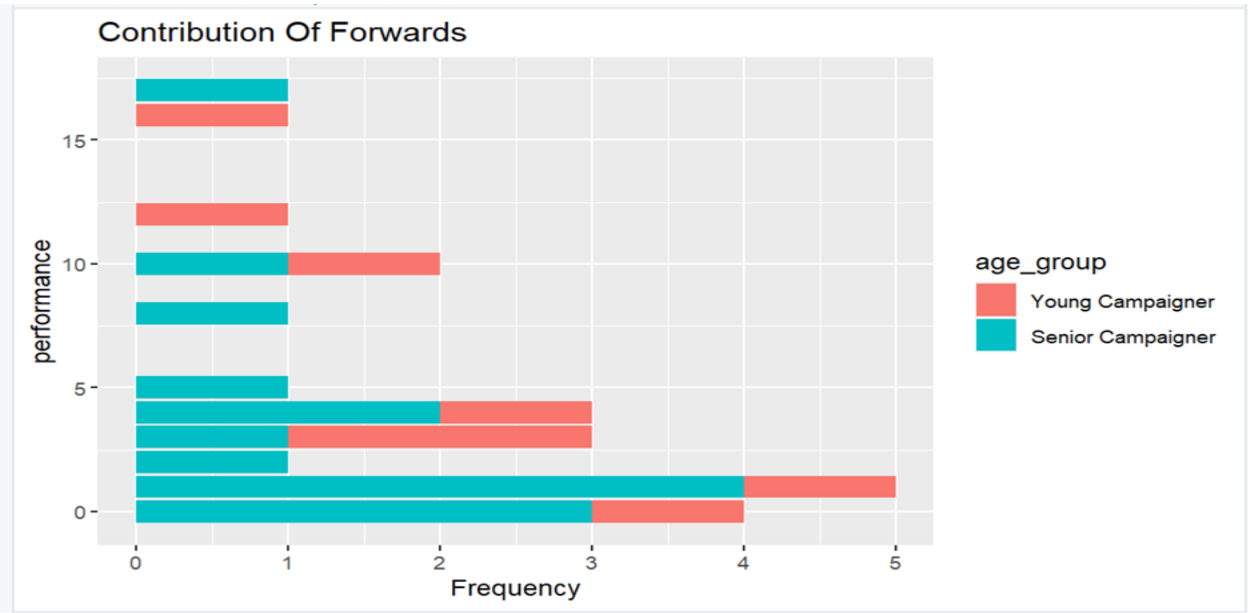


Figure 32: Contribution of the forwards for the respective team.

<u>#6) Now we run a comparison between Real Madrid's two most prolific players, Dani Carvajal and Jesús Vallejo Goals between Dani & Jesús,</u>

<u>Code:</u>

```
player1 <- result[(result$Name=="Dani Carvajal"),]
player1

player2 <- result[(result$Name=="Jesús Vallejo"),]
player2

mr <- rbind(player1,player2)
mr
g=(mr$Goal+mr$Assists)
ggplot(mr,aes(x= mr$Name, y= g, fill= mr$Name))+
 geom_bar(stat = "identity")+
 labs(x="Names",y="Goals", title = "player1 Vs player2")
```

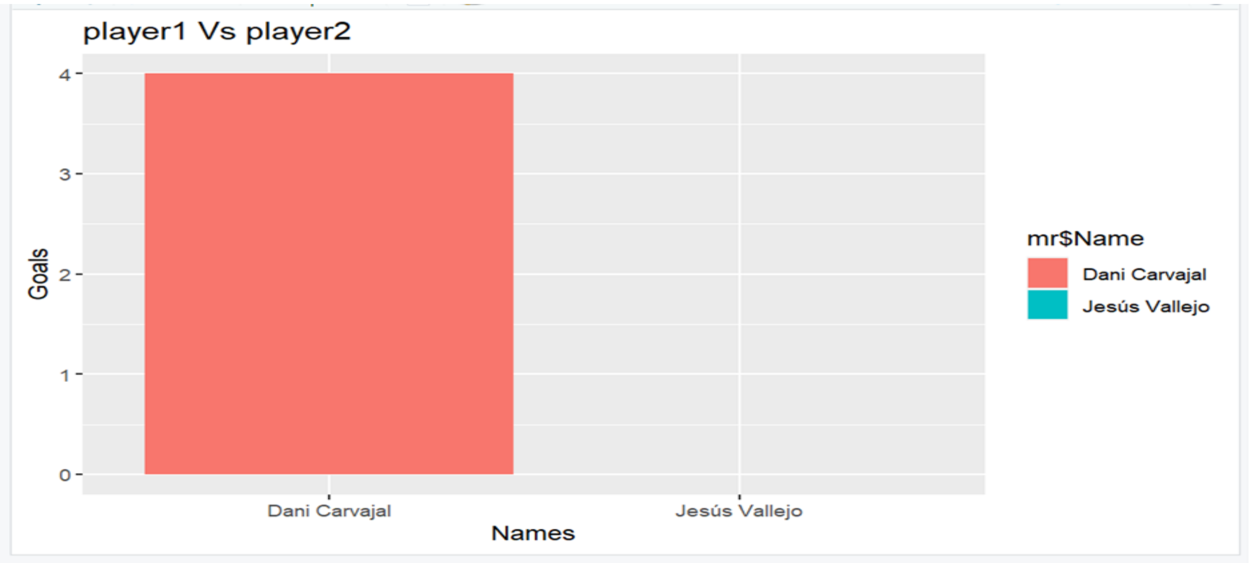


Figure 33: Real Madrid's two most prolific players, Dani Carvajal and Jesús Vallejo Goals between Dani & Jesús.

<u>#7) Now we visualize the performance of Senior and Young campaigners,</u>

<u>Code:</u>

```
ggplot(result, aes(x= age_group, fill= performace))+
  geom_bar(position = "dodge")+
  facet_wrap(~age_group)
```

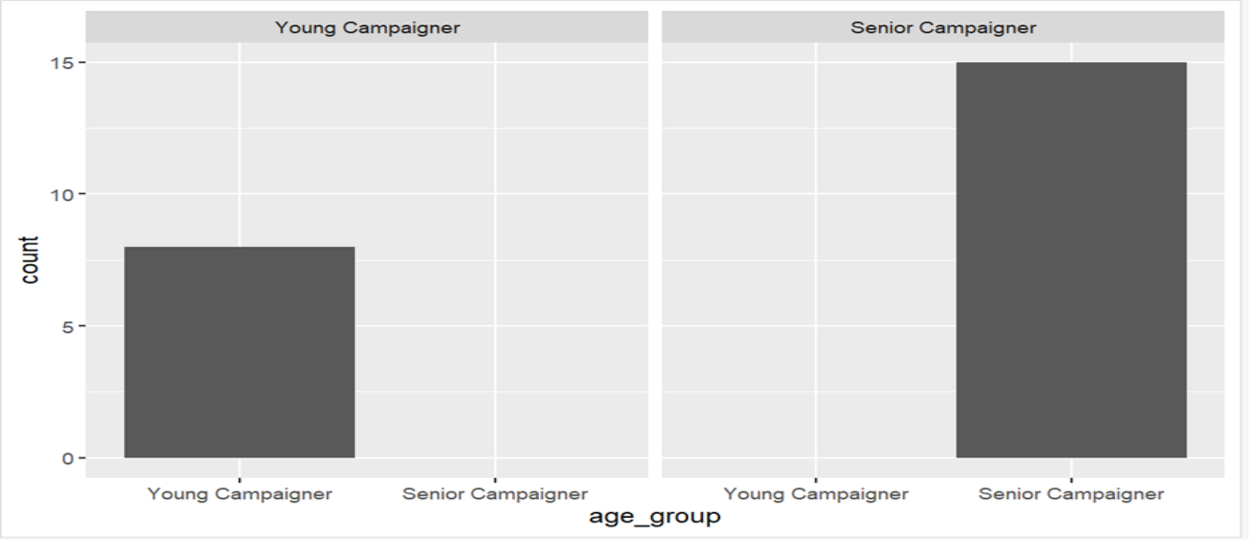


Figure 34: The performance of Senior and Young campaigners.

#8) Most fouls suffered between Messi and Ronaldo

Code:

```
barplot(mr$FA, names.arg = mr$Name)
```

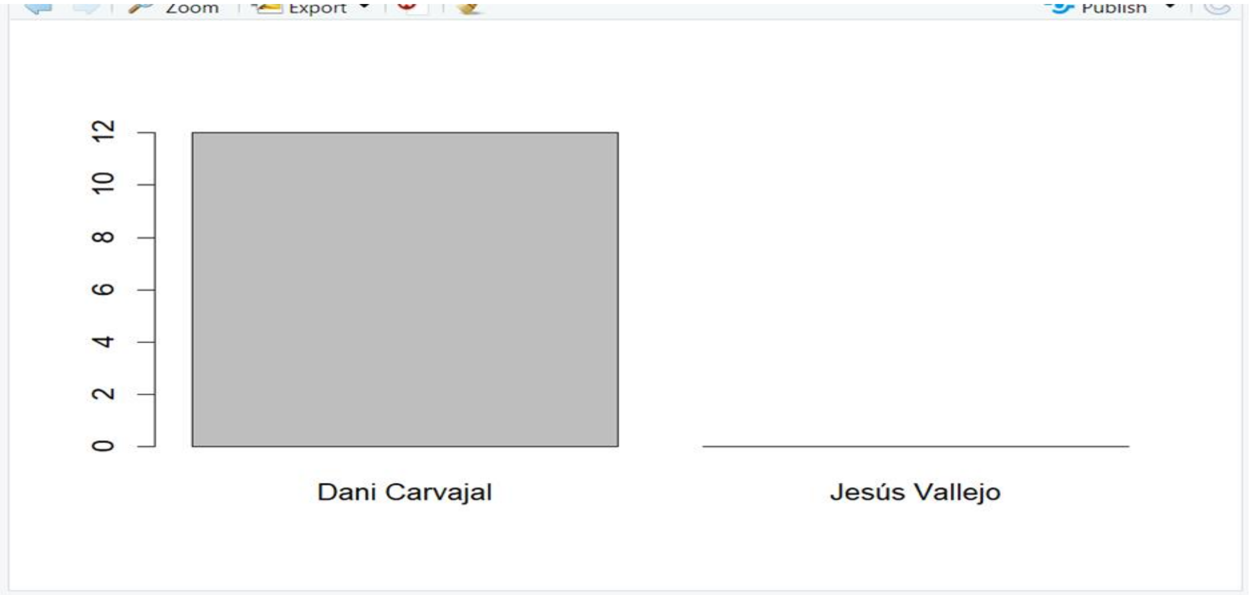


Figure 35: In this bar graph, it is clear that Dani was the most fouled player among their rivals and from the previous graphs we saw that he also had an astonishing performance, showing why he got the Balon d’or that year.

#9) We visualize the minimum height of Senior and Young campaigners’ team through a bar plot:

Code:

```
c = result$age_group
plotresult <- result %>%
group_by(result$age_group) %>%
summarise(mean=mean(HT))
View(plotresult)
plotresult<-rename(plotresult, "Tname"="result$age_group")
ggplot(plotresult, aes(x= reorder(Tname, mean), y= mean))+ geom_bar(stat="identity")+
labs(x="age_group",y="", title = "Mean Height")
```

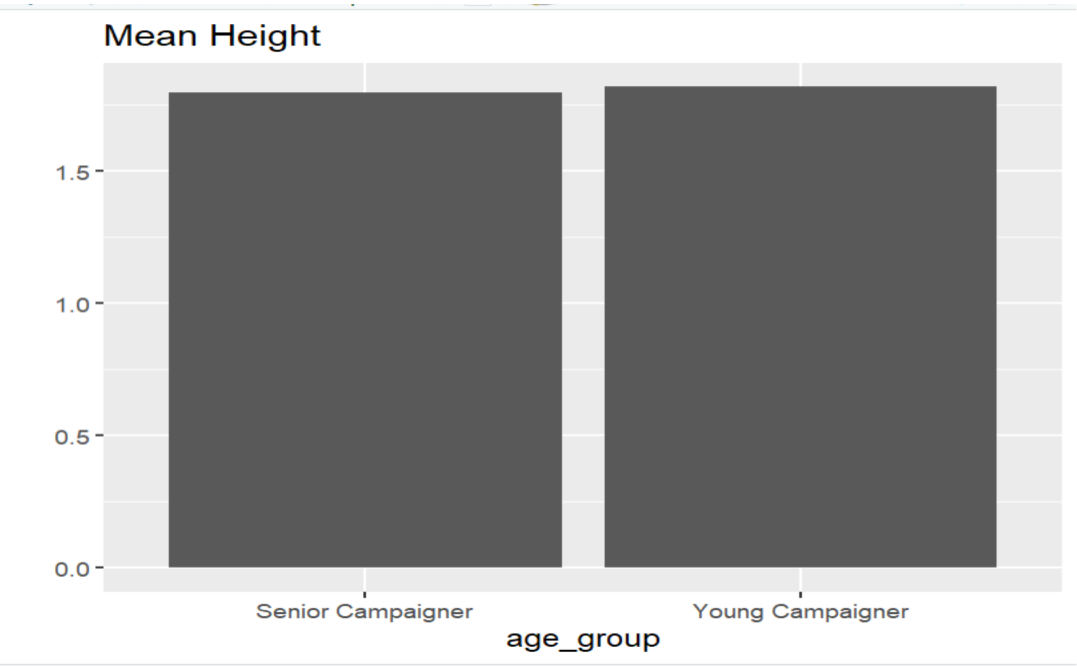
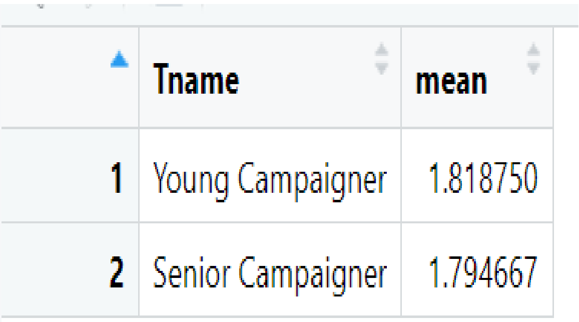


Figure 36: The minimum height of Senior and Young campaigners' team through a bar plot.



	Tname	mean
1	Young Campaigner	1.818750
2	Senior Campaigner	1.794667

Figure 37: The minimum mean height of Senior and Young campaigners' team.

#Mean Goals:

Code:

```
c = result$age_group
plotresult2 <- result %>%
 group_by(result$age_group) %>%
 summarise(mean=mean(Goal))
View(plotresult2)
```

```
plotresult2<-rename(plotresult2, "Tname"="result$age_group")
ggplot(plotresult2, aes(x= reorder(Tname, mean), y= mean))+
  geom_bar(stat="identity")
```

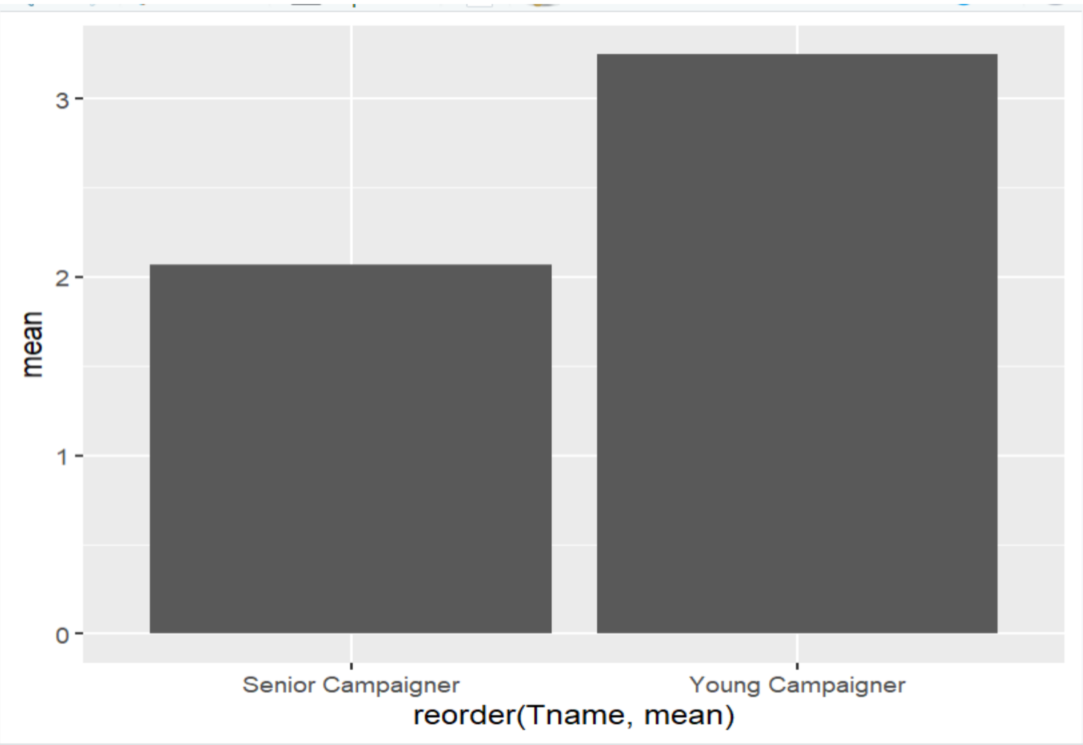
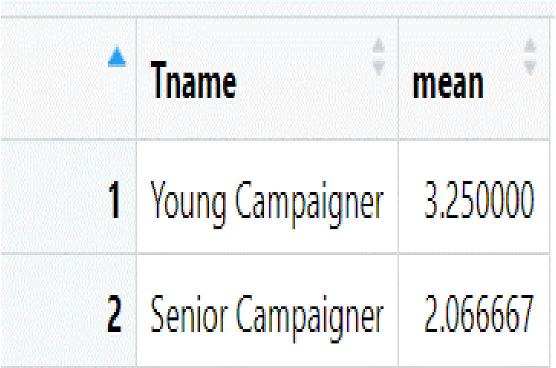


Figure 38: The minimum goals of Senior and Young campaigners' team through a bar plot.



	Tname	mean
1	Young Campaigner	3.250000
2	Senior Campaigner	2.066667

Figure 39: The minimum mean goals of Senior and Young campaigners' team.

#10) We all love players that can do both which is attack and defend. Here we try to find top goal-scoring defenders among the team:

Code:

Goal Scoring Defenders:

```
result %>% filter(result$Goal>=2 & result$POS == "D") %>%
 ggplot(aes(x= Name, y= Goal, fill=Name))+
 geom_bar(stat = "identity")+
 labs(x="Names",y="Goals", title = "Goal Scoring Defenders")
```

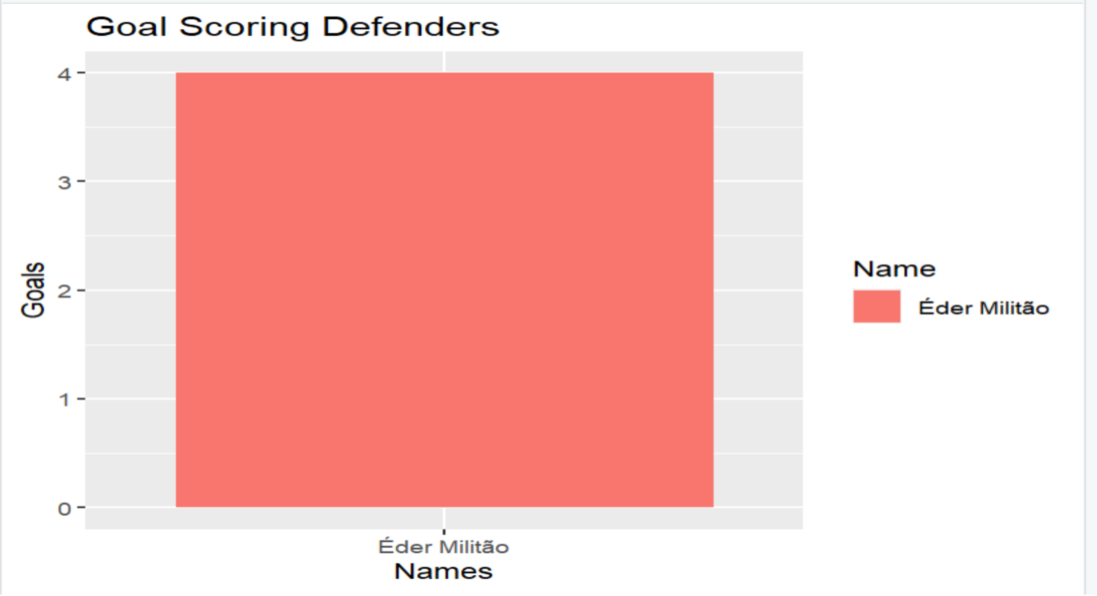


Figure 40: Top goal-scoring defenders among the team.

#11) Next, we try to figure out the greatest number of goals scored by countries:

Code:

```
result %>% ggplot(aes(x= Nation, y= Goal, fill=Nation))+
 geom_bar(stat = "identity")+
 labs(x="Nationality",y="Goals", title = "Goal Scorers By Nationality")+
 facet_wrap(~age_group)+
 coord_flip()
```

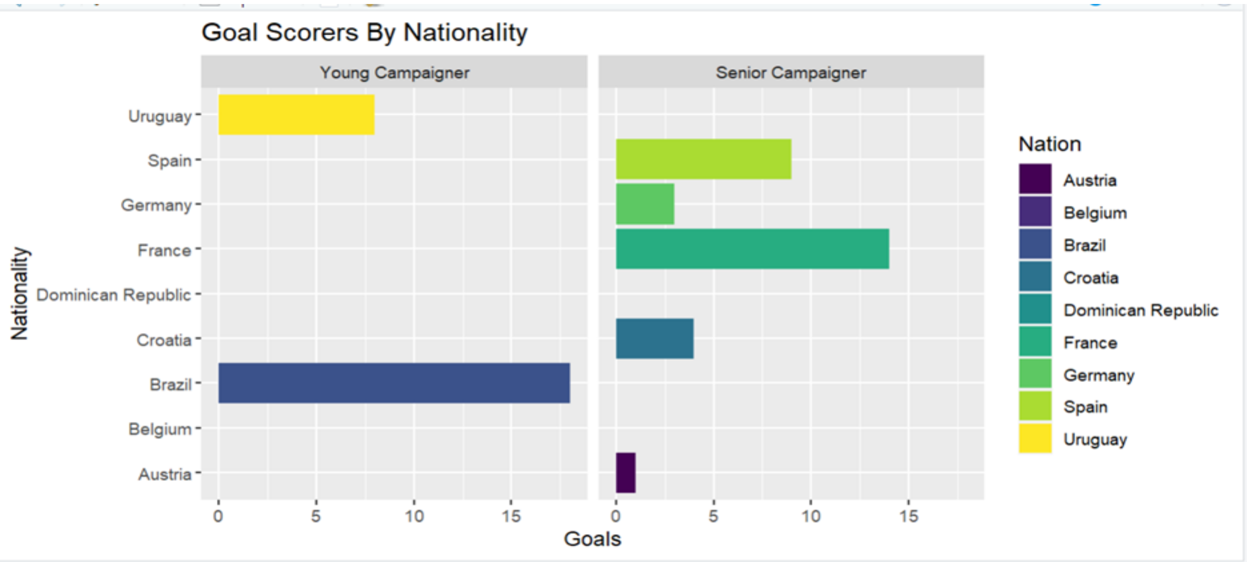


Figure 41: Goal Scores by Nationality.

#12) Now we compare the forward of the club based on goals, Forward Comparison:

Code:

```
result %>% filter(result$POS =="F" & result$Goal >mean(result$Goal)) %>%
 ggplot(aes(x= Name, y= Goal, fill=Name))+
 geom_bar(stat = "identity")+
 labs(x="Players",y="Goals", title = "Forward Comparision")+
 facet_wrap(~age_group)+
 coord_flip()
```

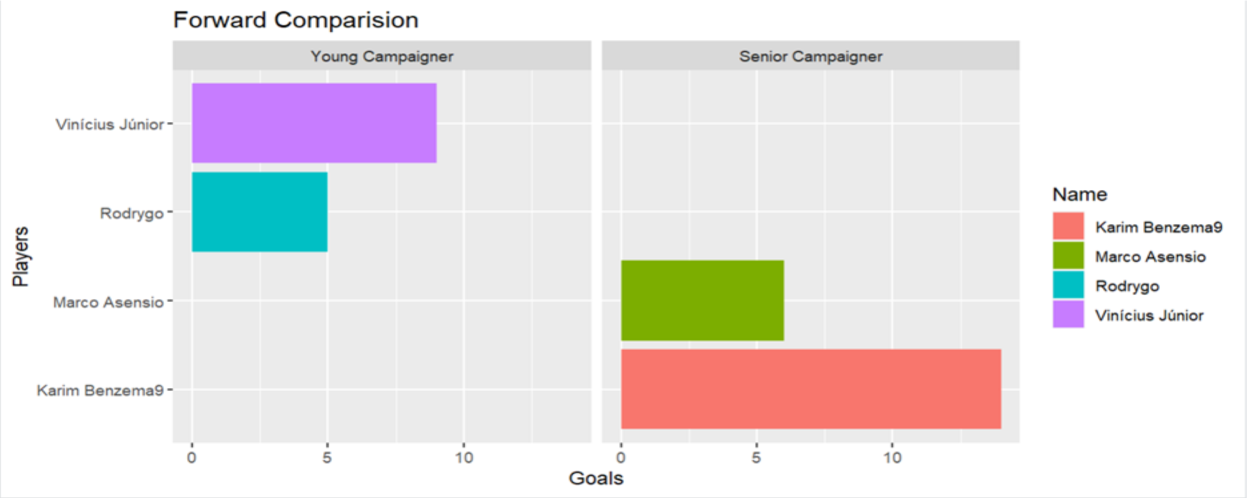


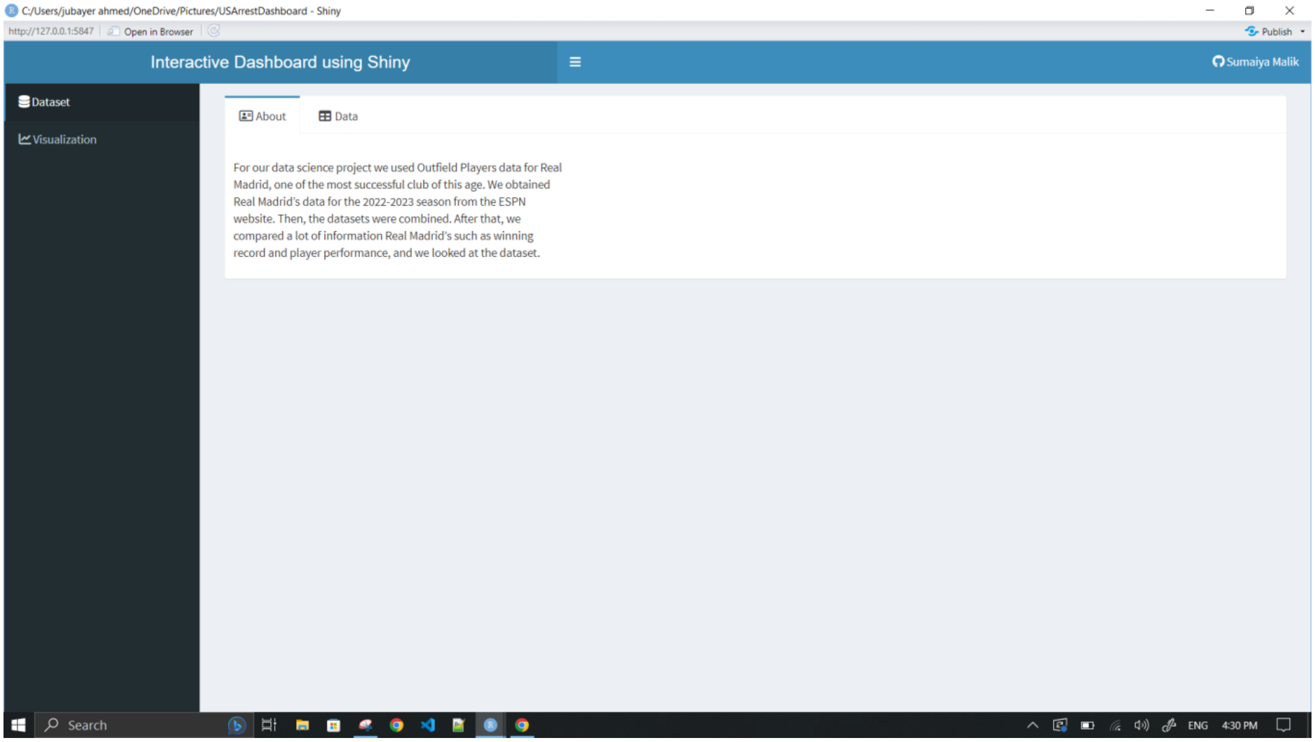
Figure 42: Forward Comparison Senior and Young campaigners.

Shiny Dashboard Implementation:

Shiny Dashboard is an open-source web framework for building interactive web applications using the R programming language. It is widely used in data science to create interactive dashboards that allow users to visualize and explore data. Shiny Dashboard Implementation in data science refers to the process of creating a Shiny Dashboard application to display and interact with data.

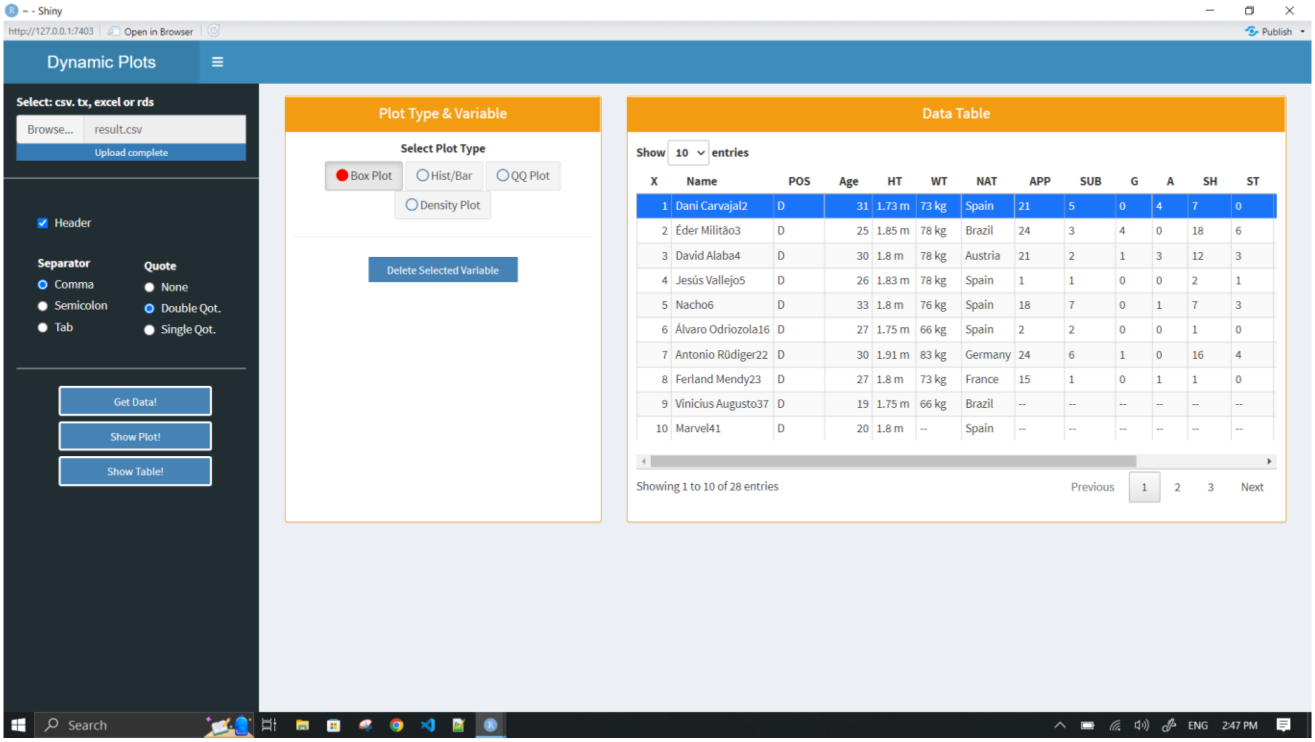
For the shiny dashboard implementation, we tried to create a reactive app based on our topic. We tried to show a reactive scatter plot and a bar plot.

We also included About, Data, Structure and Summary sections.



For our data science project we used Outfield Players data for Real Madrid, one of the most successful club of this age. We obtained Real Madrid's data for the 2022-2023 season from the ESPN website. Then, the datasets were combined. After that, we compared a lot of information Real Madrid's such as winning record and player performance, and we looked at the dataset.

Figure 43: Interactive Dashboard Using Shiny.



☑ Header

Separator: ● Comma ○ Semicolon ○ Tab

Quote: ○ None ● Double Qot. ○ Single Qot.

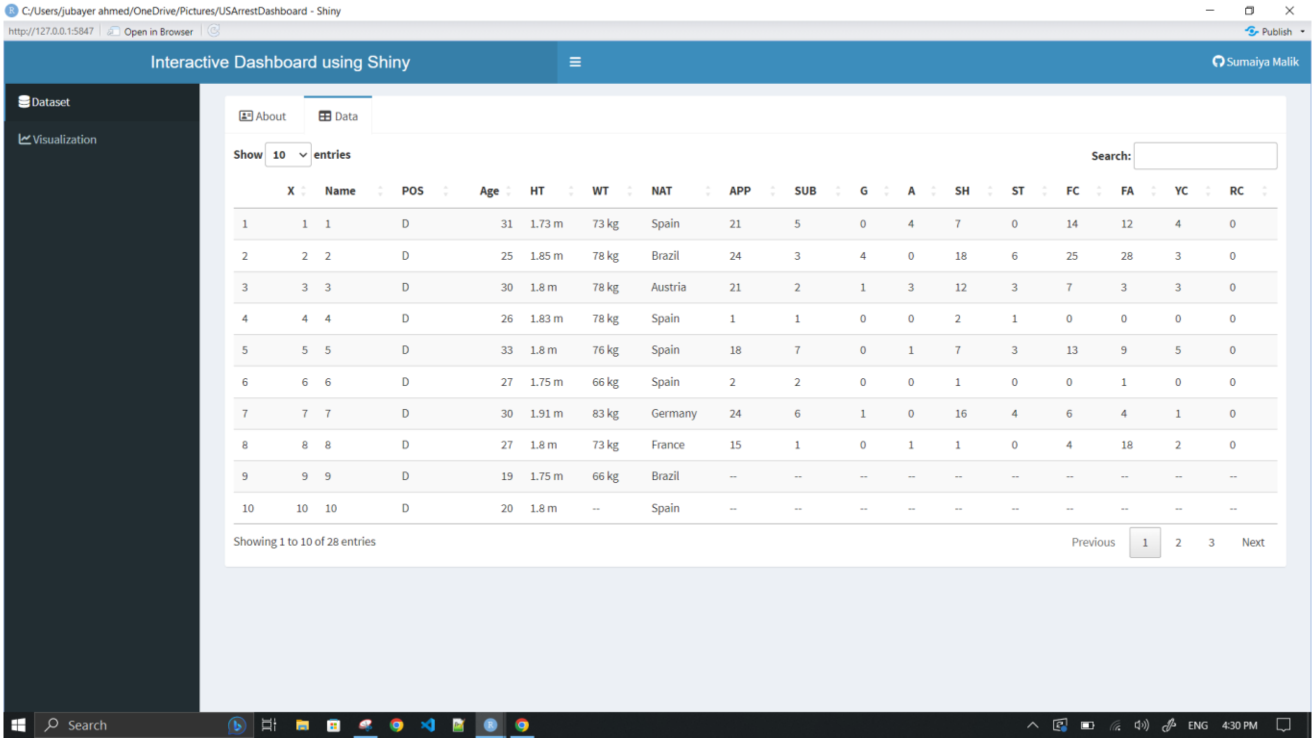
Select Plot Type: ● Box Plot ○ Hist/Bar ○ QQ Plot ○ Density Plot

Show 10 entries

X	Name	POS	Age	HT	WT	NAT	APP	SUB	G	A	SH	ST
1	Dani Carvajal2	D	31	1.73 m	73 kg	Spain	21	5	0	4	7	0
2	Éder Militão3	D	25	1.85 m	78 kg	Brazil	24	3	4	0	18	6
3	David Alaba4	D	30	1.8 m	78 kg	Austria	21	2	1	3	12	3
4	Jesús Vallejo5	D	26	1.83 m	78 kg	Spain	1	1	0	0	2	1
5	Nacho6	D	33	1.8 m	76 kg	Spain	18	7	0	1	7	3
6	Álvaro Odriozola16	D	27	1.75 m	66 kg	Spain	2	2	0	0	1	0
7	Antonio Rüdiger22	D	30	1.91 m	83 kg	Germany	24	6	1	0	16	4
8	Ferland Mendy23	D	27	1.8 m	73 kg	France	15	1	0	1	1	0
9	Vinicius Augusto37	D	19	1.75 m	66 kg	Brazil	--	--	--	--	--	--
10	Marvel41	D	20	1.8 m	--	Spain	--	--	--	--	--	--

Showing 1 to 10 of 28 entries   Previous 1 2 3 Next

Figure 44: The data table is being shown in the shiny dashboard.

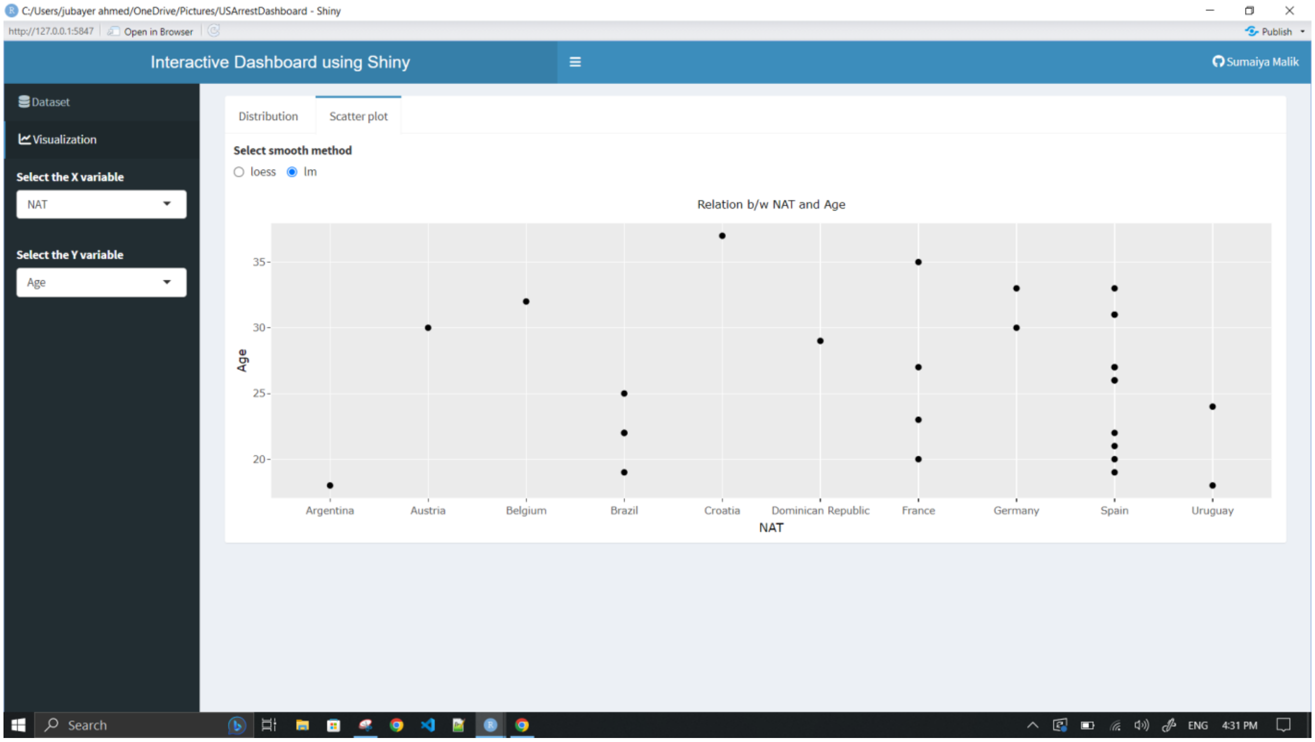


Show 10 entries   Search:

	X	Name	POS	Age	HT	WT	NAT	APP	SUB	G	A	SH	ST	FC	FA	YC	RC
1	1	1	D	31	1.73 m	73 kg	Spain	21	5	0	4	7	0	14	12	4	0
2	2	2	D	25	1.85 m	78 kg	Brazil	24	3	4	0	18	6	25	28	3	0
3	3	3	D	30	1.8 m	78 kg	Austria	21	2	1	3	12	3	7	3	3	0
4	4	4	D	26	1.83 m	78 kg	Spain	1	1	0	0	2	1	0	0	0	0
5	5	5	D	33	1.8 m	76 kg	Spain	18	7	0	1	7	3	13	9	5	0
6	6	6	D	27	1.75 m	66 kg	Spain	2	2	0	0	1	0	0	1	0	0
7	7	7	D	30	1.91 m	83 kg	Germany	24	6	1	0	16	4	6	4	1	0
8	8	8	D	27	1.8 m	73 kg	France	15	1	0	1	1	0	4	18	2	0
9	9	9	D	19	1.75 m	66 kg	Brazil	--	--	--	--	--	--	--	--	--	--
10	10	10	D	20	1.8 m	--	Spain	--	--	--	--	--	--	--	--	--	--

Showing 1 to 10 of 28 entries   Previous 1 2 3 Next

Figure 45: The total data frame after applying Pre-processing on shiny.



C:/Users/jubayer ahmed/OneDrive/Pictures/USArrestDashboard - Shiny

http://127.0.0.1:5847 | Open in Browser

Interactive Dashboard using Shiny

Select smooth method

○ loess ● lm

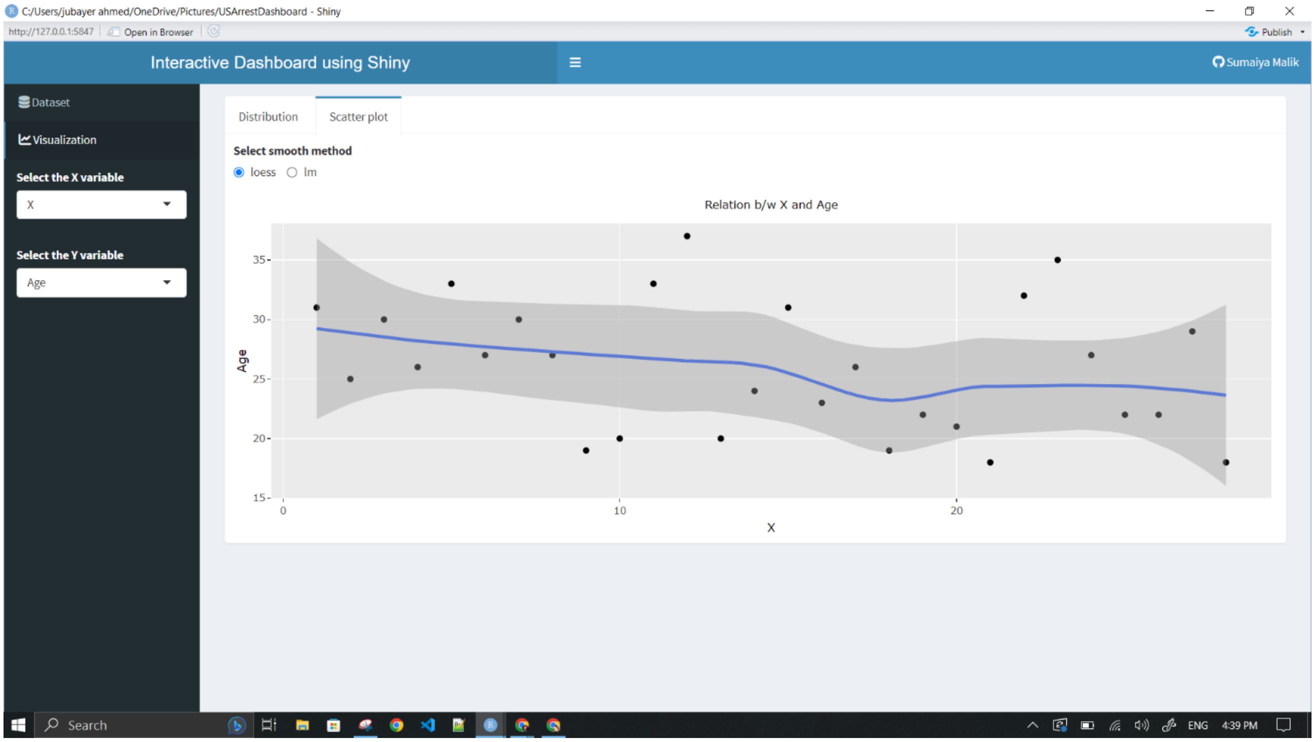
Select the X variable

NAT

Select the Y variable

Age

Figure 46: Scatter plot obtained from data table is being shown in the shiny dashboard.



C:/Users/jubayer ahmed/OneDrive/Pictures/USArrestDashboard - Shiny

http://127.0.0.1:5847 | Open in Browser

Interactive Dashboard using Shiny

Select smooth method

● loess ○ lm

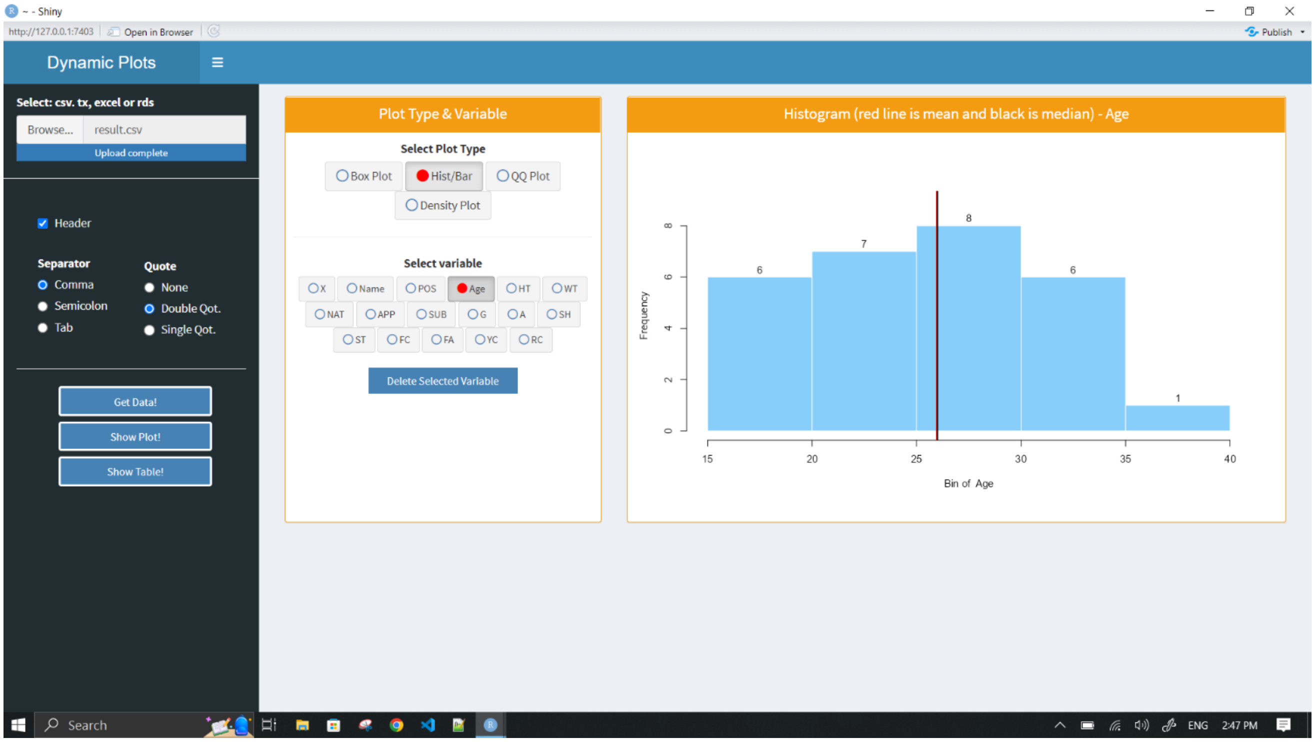
Select the X variable

X

Select the Y variable

Age

Figure 47: Scatter plot with regression line obtained from data table is being shown in the shiny dashboard.



~ - Shiny

http://127.0.0.1:7403 | Open in Browser

Dynamic Plots

Select: csv. tx, excel or rds

Browse... result.csv

Upload complete

☑ Header

Separator: ● Comma ○ Semicolon ○ Tab

Quote: ○ None ● Double Qot. ○ Single Qot.

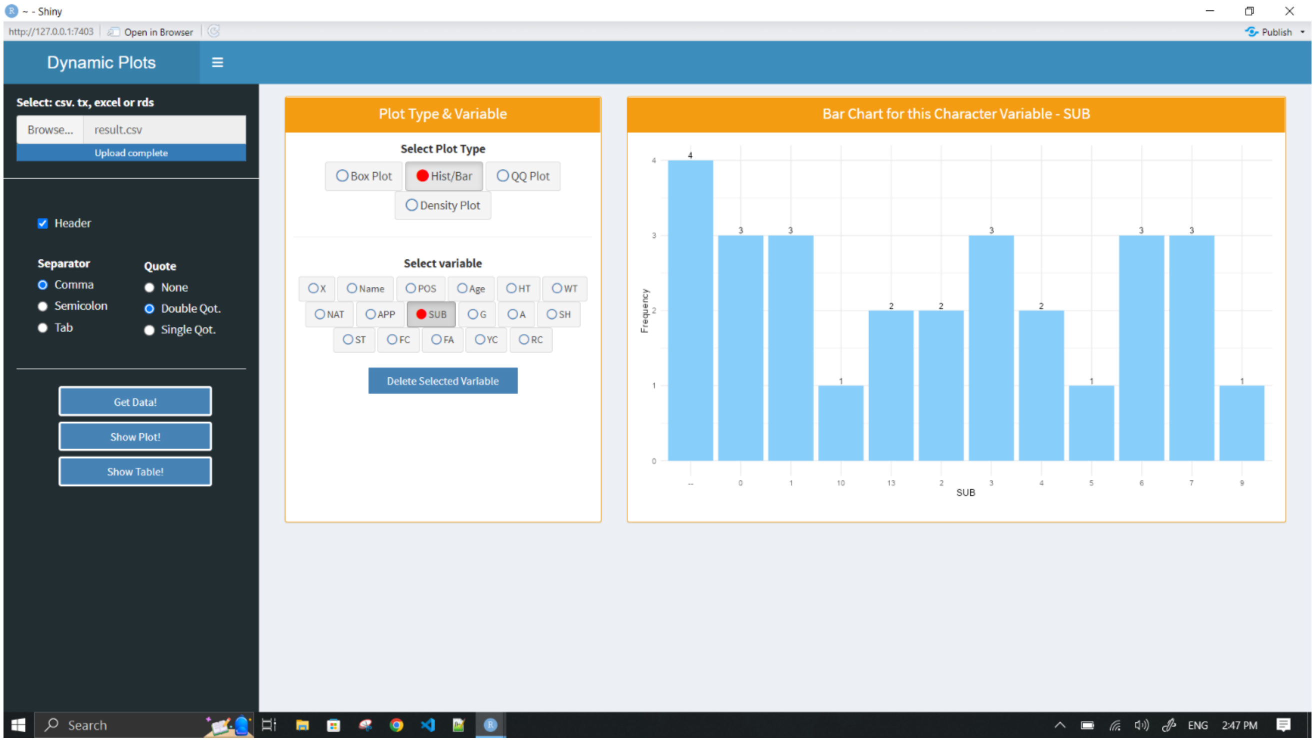
Plot Type & Variable

Select Plot Type: ○ Box Plot ● Hist/Bar ○ QQ Plot ○ Density Plot

Select variable: ○ X ○ Name ○ POS ● Age ○ HT ○ WT ○ NAT ○ APP ○ SUB ○ G ○ A ○ SH ○ ST ○ FC ○ FA ○ YC ○ RC

Histogram (red line is mean and black is median) - Age

Figure 48: The Histogram obtained from data table is being shown in the shiny dashboard.



~ - Shiny

http://127.0.0.1:7403 | Open in Browser

Dynamic Plots

Select: csv. tx, excel or rds

Browse... result.csv

Upload complete

☑ Header

Separator: ● Comma ○ Semicolon ○ Tab

Quote: ○ None ● Double Qot. ○ Single Qot.

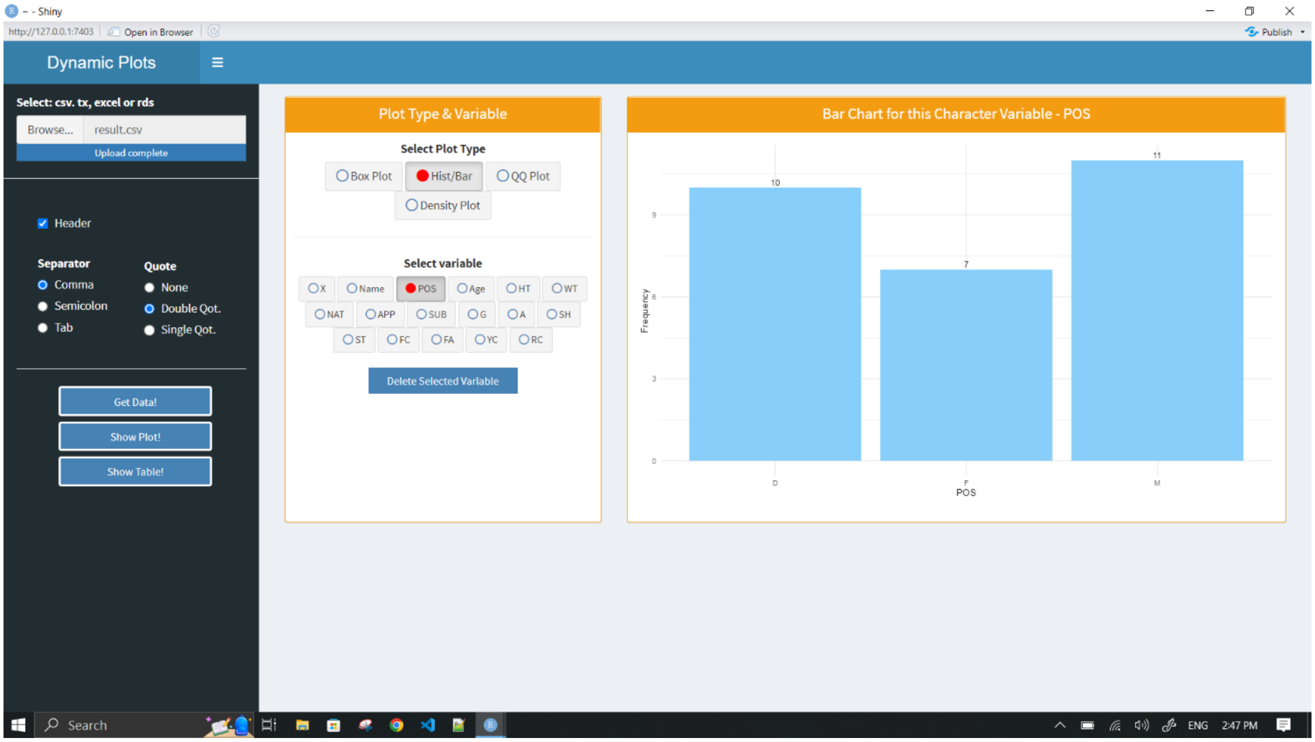
Plot Type & Variable

Select Plot Type: ○ Box Plot ● Hist/Bar ○ QQ Plot ○ Density Plot

Select variable: ○ X ○ Name ○ POS ○ Age ○ HT ○ WT ○ NAT ○ APP ● SUB ○ G ○ A ○ SH ○ ST ○ FC ○ FA ○ YC ○ RC

Bar Chart for this Character Variable - SUB

Figure 49: The Bar Chart obtained from data table is being shown in the shiny dashboard.



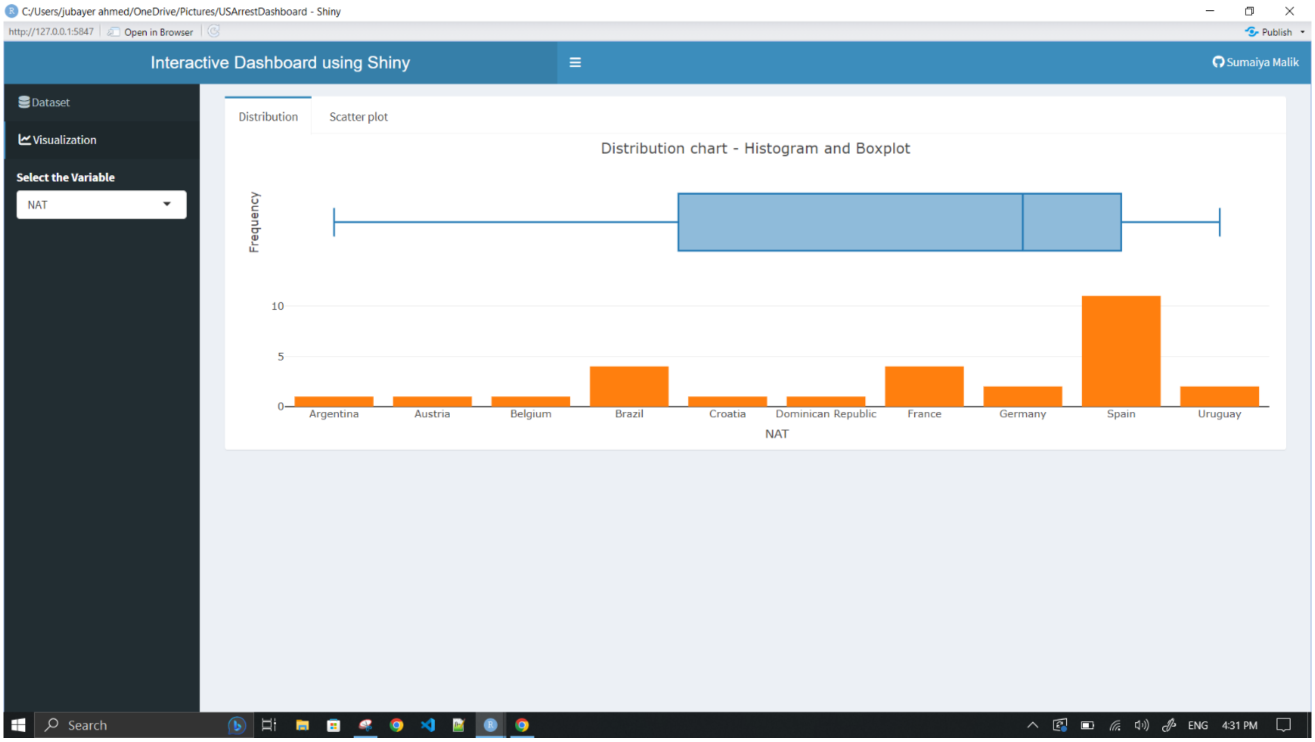
result.csv

☑ Header

Separator: ● Comma ○ Semicolon ○ Tab

Quote: ○ None ● Double Qot. ○ Single Qot.

Figure 50: The Histogram obtained from data table is being shown in the shiny dashboard.



Select the Variable: NAT

Figure 51: Quartile values of variables.