CIS 5250 – 01 VISUAL ANALYTICS

R-Project

126 Years of Historical Olympic Dataset

Sirisha Mahesh, Spencer Asahi

Introduction and Data set URL’s

This dataset actually provides an extensive view of athletes, their characteristics, and their performance in the Olympic Games, it gives certain points for in-depth analysis and insights into trends, relationships, and outcomes. Personal data such as gender, height, weight, and date of birth provide demographic and physical profiles of athletes, which can be used to study trends across different sports, genders, or countries. The inclusion of country information, along with National Olympic Committee (NOC) codes, further enables the cross examination of regional or national strengths in specific disciplines. Additional attributes like special notes and descriptions offer a deeper context into the personal achievements or backgrounds of athletes, enriching the narrative and origins behind their performances. Because all athletes are made differently.

Event-specific details, such as the sport, competition, position, medal type, and the number of participants, capture detailed performance metrics, enabling comparisons across events, athletes, and the different editions of the Olympics. Information about result dates, locations, and formats adds layers of context to these performances, allowing for analysis of how different factors, such as venue or competition structure, may influence certain types of outcomes. Such as the growth of certain sports for further examination. Furthermore, the dataset also includes essential contextual details about the Olympic Games themselves, such as the year, host city, start and end dates, and the overall length of the competition period. These attributes enable analysts to link individual performances and event data to the historical and geographical context of each Olympic edition.

So by pretty much combining data on athletes, events, and Olympic editions, this dataset facilitates the exploration of relationships between personal attributes, event outcomes, and the broader dynamics of the games. For example, correlations can be drawn between athlete height and success in certain sports or between the number of participants in an event and medal outcomes. Tracking medal distributions across editions and countries provides insights into the evolution of national performance over time, while studying the host country's impact on results highlights potential advantages of hosting. It's like playing on your own turf. In summary, this dataset serves as a powerful tool for understanding the intricate relationships between athletes, their performances, and the larger context of the Olympic Games, enabling rich analysis of trends and outcomes across time and geography.

References:

1.https://www.kaggle.com/datasets/muhammadehsan02/126-years-of-historical-olympic-dataset

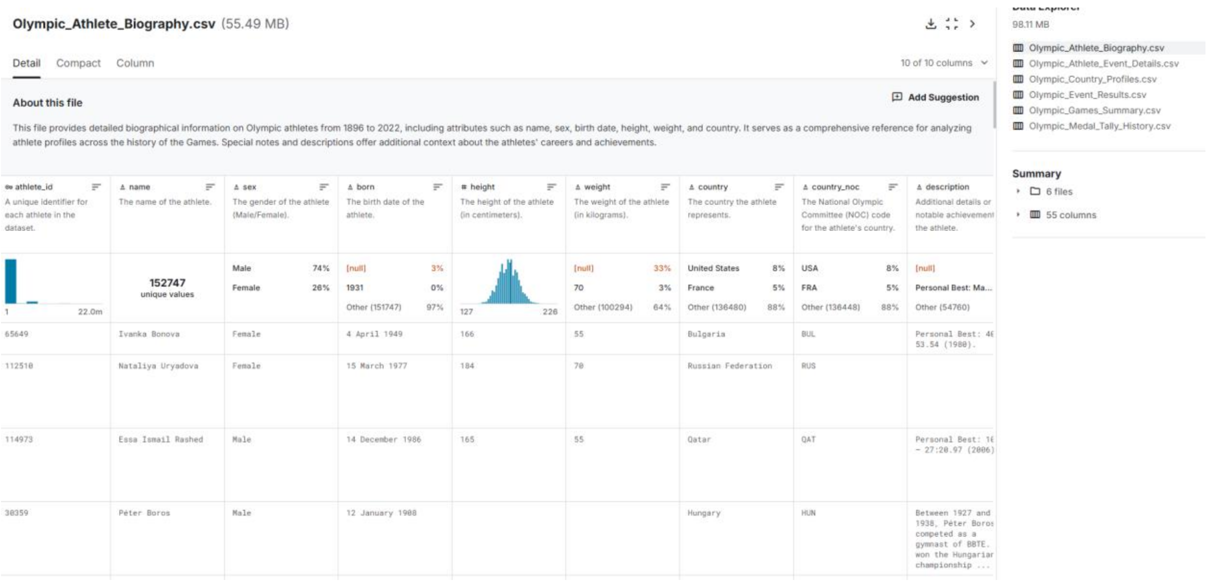
2 .Apa formatted this

McMahon, J., & Penney, D. (2023). *Exploring athletes' well-being: Psychological and contextual factors in sports performance. CISS Journal*

3. https://ciss-journal.org/article/view/9363

Data Description:

Column Name	Description
Athlete id	A Unique identifier for each athlete in the dataset
Name	The name of the athlete
Sex	The gender of the athlete(Male/Female)
Height	The height of the athlete in centimeters.
Weight	The weight of the athlete in kilograms
Country	The National Olympic Committee (NOC) code for the athlete's Country.
Edition	The year or edition of the Olympic Games
Sport	The sport in which the athlete completed.
Event	The specific event or competition within the sport.
Result id	A unique identifier for the athlete's result
isTeamSport	Identify whether it's a team Sport or individual
Medal	The Type of medal awarded (Gold, Silver, Bronze or None)
Result Participants	The number of participants in the events.
Result format	The format of the event result.



Olympic_Athlete_Biography.csv (55.49 MB)

Detail Compact Column

10 of 10 columns

About this file

Add Suggestion

This file provides detailed biographical information on Olympic athletes from 1896 to 2022, including attributes such as name, sex, birth date, height, weight, and country. It serves as a comprehensive reference for analyzing athlete profiles across the history of the Games. Special notes and descriptions offer additional context about the athletes' careers and achievements.

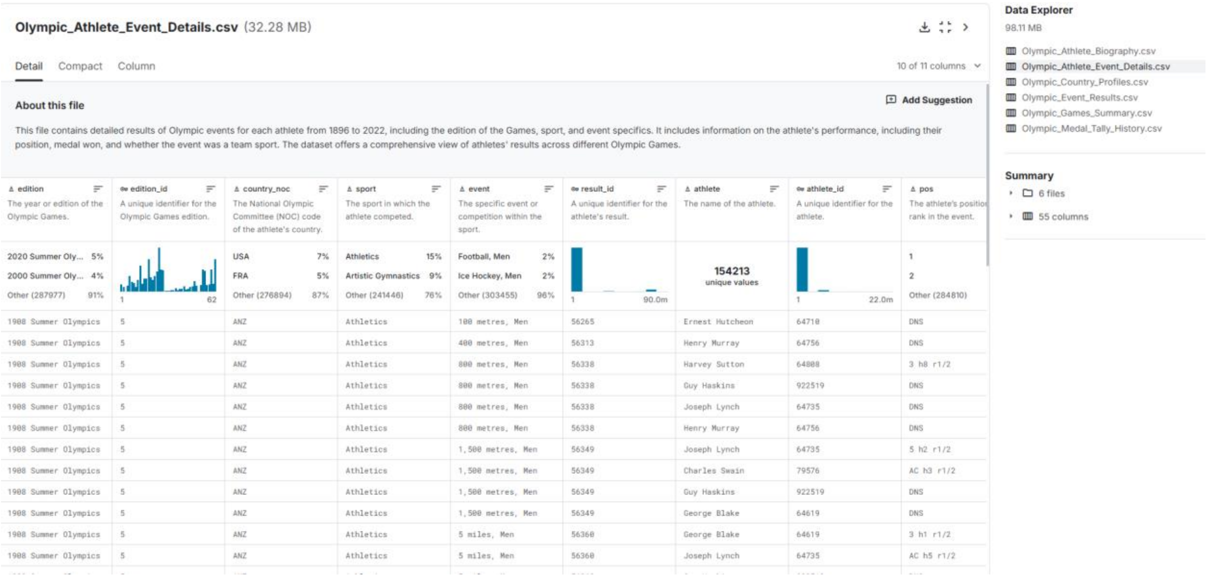
athlete_id	name	sex	born	height	weight	country	country_noc	description
A unique identifier for each athlete in the dataset.	The name of the athlete.	The gender of the athlete (Male/Female).	The birth date of the athlete.	The height of the athlete (in centimeters).	The weight of the athlete (in kilograms).	The country the athlete represents.	The National Olympic Committee (NOC) code for the athlete's country.	Additional details or notable achievement the athlete.
1 22.0m	152747 unique values	Male 74% Female 26%	[null] 3% 1931 0% Other (151747) 97%	127 226	[null] 33% 70 3% Other (100294) 64%	United States 8% France 5% Other (136480) 88%	USA 8% FRA 5% Other (136448) 88%	[null] Personal Best: Ma... Other (54760)
65649	Ivanka Bonova	Female	4 April 1949	166	55	Bulgaria	BUL	Personal Best: 4€ 53.54 (1980).
112510	Nataliya Uryadova	Female	15 March 1977	184	70	Russian Federation	RUS	
114973	Essa Ismail Rashed	Male	14 December 1986	165	55	Qatar	QAT	Personal Best: 1€ – 27:20.97 (2006)
38359	Péter Boros	Male	12 January 1908			Hungary	HUN	Between 1927 and 1938, Péter Boros competed as a gymnast of BBTE. won the Hungariar championship ...

98.11 MB

- Olympic_Athlete_Biography.csv
- Olympic_Athlete_Event_Details.csv
- Olympic_Country_Profiles.csv
- Olympic_Event_Results.csv
- Olympic_Games_Summary.csv
- Olympic_Medal_Tally_History.csv

Summary

- 6 files
- 55 columns



Olympic_Athlete_Event_Details.csv (32.28 MB)

Detail Compact Column 10 of 11 columns

About this file Add Suggestion

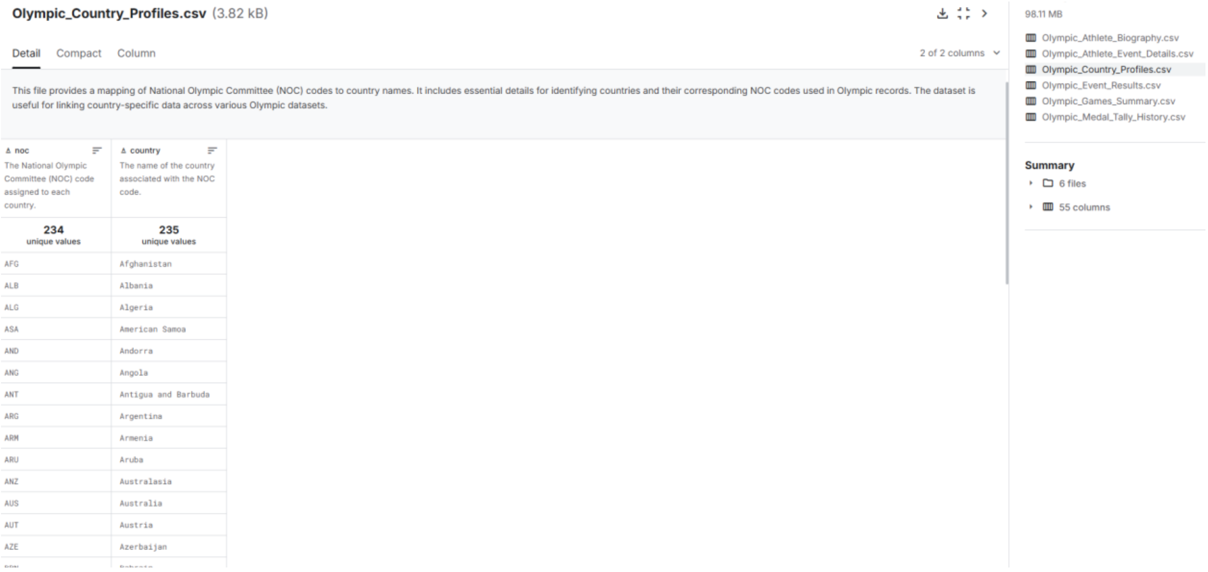
This file contains detailed results of Olympic events for each athlete from 1896 to 2022, including the edition of the Games, sport, and event specifics. It includes information on the athlete's performance, including their position, medal won, and whether the event was a team sport. The dataset offers a comprehensive view of athletes' results across different Olympic Games.

edition	edition_id	country_noc	sport	event	result_id	athlete	athlete_id	pos
The year or edition of the Olympic Games.	A unique identifier for the Olympic Games edition.	The National Olympic Committee (NOC) code of the athlete's country.	The sport in which the athlete competed.	The specific event or competition within the sport.	A unique identifier for the athlete's result.	The name of the athlete.	A unique identifier for the athlete.	The athlete's positio rank in the event.
2020 Summer Oly... 5% 2000 Summer Oly... 4% Other (287977) 91%	1 62	USA 7% FRA 5% Other (276894) 87%	Athletics 15% Artistic Gymnastics 9% Other (241446) 76%	Football, Men 2% Ice Hockey, Men 2% Other (303455) 96%	1 90.0m	154213 unique values	1 22.0m	1 2 Other (284810)
1908 Summer Olympics	5	ANZ	Athletics	100 metres, Men	56265	Ernest Hutcheon	64710	DNS
1908 Summer Olympics	5	ANZ	Athletics	400 metres, Men	56313	Henry Murray	64756	DNS
1908 Summer Olympics	5	ANZ	Athletics	800 metres, Men	56338	Harvey Sutton	64808	3 h8 r1/2
1908 Summer Olympics	5	ANZ	Athletics	800 metres, Men	56338	Guy Haskins	922519	DNS
1908 Summer Olympics	5	ANZ	Athletics	800 metres, Men	56338	Joseph Lynch	64735	DNS
1908 Summer Olympics	5	ANZ	Athletics	800 metres, Men	56338	Henry Murray	64756	DNS
1908 Summer Olympics	5	ANZ	Athletics	1,500 metres, Men	56349	Joseph Lynch	64735	5 h2 r1/2
1908 Summer Olympics	5	ANZ	Athletics	1,500 metres, Men	56349	Charles Swain	79576	AC h3 r1/2
1908 Summer Olympics	5	ANZ	Athletics	1,500 metres, Men	56349	Guy Haskins	922519	DNS
1908 Summer Olympics	5	ANZ	Athletics	1,500 metres, Men	56349	George Blake	64619	DNS
1908 Summer Olympics	5	ANZ	Athletics	5 miles, Men	56360	George Blake	64619	3 h1 r1/2
1908 Summer Olympics	5	ANZ	Athletics	5 miles, Men	56360	Joseph Lynch	64735	AC h5 r1/2

Data Explorer
98.11 MB

- Olympic_Athlete_Biography.csv
- Olympic_Athlete_Event_Details.csv
- Olympic_Country_Profiles.csv
- Olympic_Event_Results.csv
- Olympic_Games_Summary.csv
- Olympic_Medal_Tally_History.csv

Summary
- 6 files
- 55 columns



Olympic_Country_Profiles.csv (3.82 kB)

Detail Compact Column 2 of 2 columns

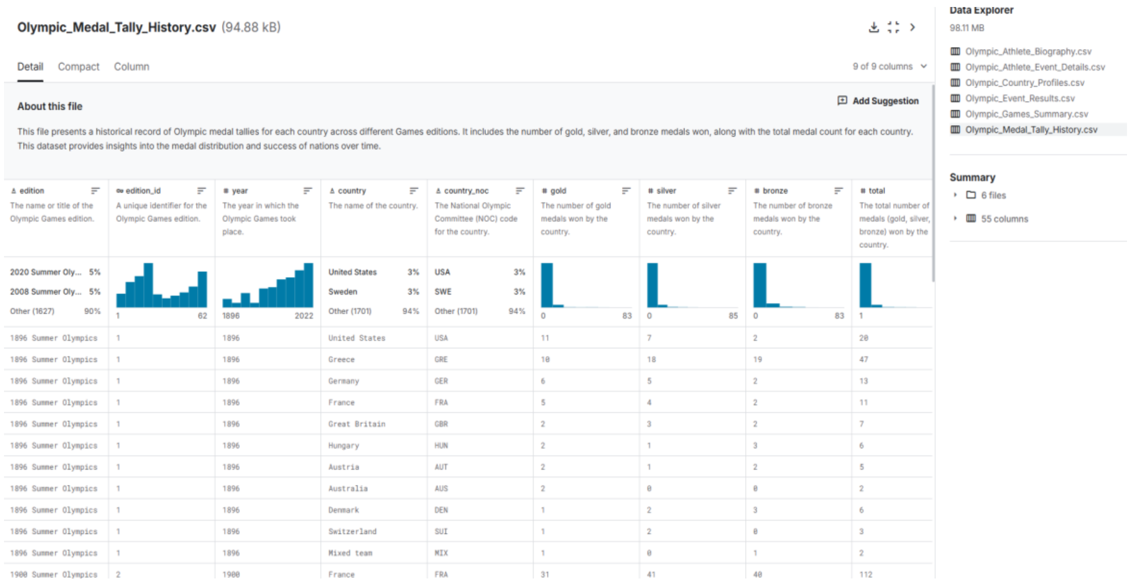
This file provides a mapping of National Olympic Committee (NOC) codes to country names. It includes essential details for identifying countries and their corresponding NOC codes used in Olympic records. The dataset is useful for linking country-specific data across various Olympic datasets.

noc	country
The National Olympic Committee (NOC) code assigned to each country.	The name of the country associated with the NOC code.
234 unique values	235 unique values
AFG	Afghanistan
ALB	Albania
ALG	Algeria
ASA	American Samoa
AND	Andorra
ANG	Angola
ANT	Antigua and Barbuda
ARG	Argentina
ARM	Armenia
ARU	Aruba
ANZ	Australasia
AUS	Australia
AUT	Austria
AZE	Azerbaijan

98.11 MB

- Olympic_Athlete_Biography.csv
- Olympic_Athlete_Event_Details.csv
- Olympic_Country_Profiles.csv
- Olympic_Event_Results.csv
- Olympic_Games_Summary.csv
- Olympic_Medal_Tally_History.csv

Summary
- 6 files
- 55 columns



Olympic_Medal_Tally_History.csv (94.88 kB)

Detail Compact Column 9 of 9 columns

About this file Add Suggestion

This file presents a historical record of Olympic medal tallies for each country across different Games editions. It includes the number of gold, silver, and bronze medals won, along with the total medal count for each country. This dataset provides insights into the medal distribution and success of nations over time.

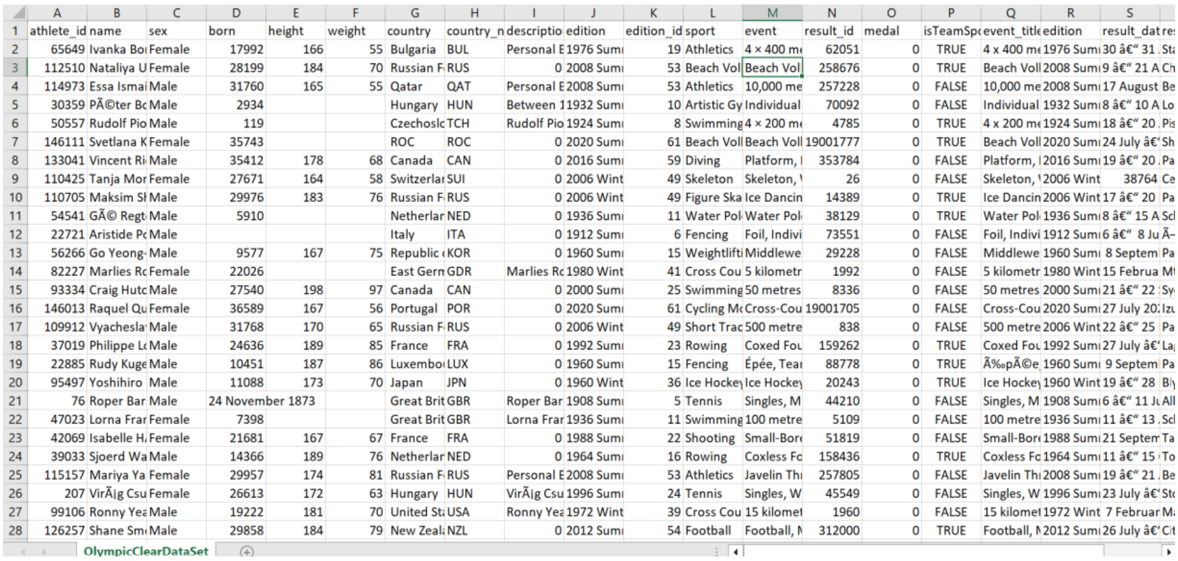
edition	edition_id	year	country	country_noc	gold	silver	bronze	total
The name or title of the Olympic Games edition.	A unique identifier for the Olympic Games edition.	The year in which the Olympic Games took place.	The name of the country.	The National Olympic Committee (NOC) code for the country.	The number of gold medals won by the country.	The number of silver medals won by the country.	The number of bronze medals won by the country.	The total number of medals (gold, silver, bronze) won by the country.
2020 Summer Oly... 5% 2008 Summer Oly... 5% Other (1627) 90%	1 62	1896 2022	United States 3% Sweden 3% Other (1701) 94%	USA 3% SWE 3% Other (1701) 94%	0 83	0 85	0 83	1
1896 Summer Olympics	1	1896	United States	USA	11	7	2	20
1896 Summer Olympics	1	1896	Greece	GRE	10	18	19	47
1896 Summer Olympics	1	1896	Germany	GER	6	5	2	13
1896 Summer Olympics	1	1896	France	FRA	5	4	2	11
1896 Summer Olympics	1	1896	Great Britain	GBR	2	3	2	7
1896 Summer Olympics	1	1896	Hungary	HUN	2	1	3	6
1896 Summer Olympics	1	1896	Austria	AUT	2	1	2	5
1896 Summer Olympics	1	1896	Australia	AUS	2	0	0	2
1896 Summer Olympics	1	1896	Denmark	DEN	1	2	3	6
1896 Summer Olympics	1	1896	Switzerland	SUI	1	2	0	3
1896 Summer Olympics	1	1896	Mixed team	MIX	1	0	1	2
1900 Summer Olympics	2	1900	France	FRA	31	41	40	112

Data Explorer
98.11 MB

- Olympic_Athlete_Biography.csv
- Olympic_Athlete_Event_Details.csv
- Olympic_Country_Profiles.csv
- Olympic_Event_Results.csv
- Olympic_Games_Summary.csv
- Olympic_Medal_Tally_History.csv

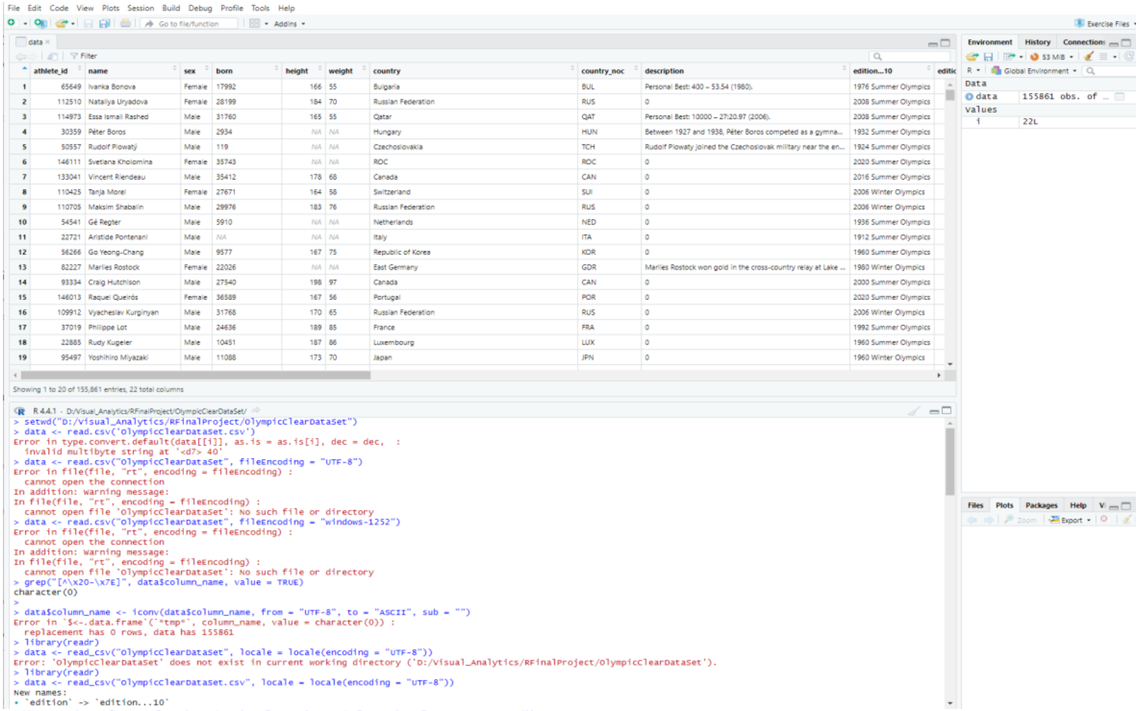
Summary
- 6 files
- 55 columns

Excel Screenshot



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	athlete_id	name	sex	born	height	weight	country	country_n	descriptio	edition	edition_id	sport	event	result_id	medal	isTeamSp	event_title	edition	result_dat	re
2	65649	Ivanka Bo	Female	17992	166	55	Bulgaria	BUL	Personal E	1976 Sum	19	Athletics	4 × 400 m	62051	0	TRUE	4 x 400 m	1976 Sum	30 â€“ 31 .	Sta
3	112510	Nataliya U	Female	28199	184	70	Russian F	RUS	0	2008 Sum	53	Beach Vol	Beach Vol	258676	0	TRUE	Beach Voll	2008 Sum	9 â€“ 21 A	Ch
4	114973	Essa Ismai	Male	31760	165	55	Qatar	QAT	Personal E	2008 Sum	53	Athletics	10,000 me	257228	0	FALSE	10,000 me	2008 Sum	17 August	Be
5	30359	PÃ©ter Bc	Male	2934			Hungary	HUN	Between 1	1932 Sum	10	Artistic Gy	Individual	70092	0	FALSE	Individual	1932 Sum	8 â€“ 10 A	Lo
6	50557	Rudolf Pio	Male	119			Czechoslc	TCH	Rudolf Pio	1924 Sum	8	Swimming	4 × 200 m	4785	0	TRUE	4 x 200 m	1924 Sum	18 â€“ 20 .	Pis
7	146111	Svetlana K	Female	35743			ROC	ROC	0	2020 Sum	61	Beach Voll	Beach Voll	19001777	0	TRUE	Beach Voll	2020 Sum	24 July â€'	Sh
8	133041	Vincent Ri	Male	35412	178	68	Canada	CAN	0	2016 Sum	59	Diving	Platform, I	353784	0	FALSE	Platform, I	2016 Sum	19 â€“ 20 .	Pa
9	110425	Tanja Mor	Female	27671	164	58	Switzerlar	SUI	0	2006 Wint	49	Skeleton	Skeleton, \	26	0	FALSE	Skeleton, \	2006 Wint	38764	Ce
10	110705	Maksim Sh	Male	29976	183	76	Russian F	RUS	0	2006 Wint	49	Figure Ska	Ice Dancin	14389	0	TRUE	Ice Dancin	2006 Wint	17 â€“ 20	Pa
11	54541	GÃ© Regt	Male	5910			Netherlar	NED	0	1936 Sum	11	Water Pol	Water Pol	38129	0	TRUE	Water Pol	1936 Sum	8 â€“ 15 A	Scl
12	22721	Aristide Pc	Male				Italy	ITA	0	1912 Sum	6	Fencing	Foil, Indivi	73551	0	FALSE	Foil, Indivi	1912 Sum	6 â€“ 8 Ju	Ã–
13	56266	Go Yeong-	Male	9577	167	75	Republic (	KOR	0	1960 Sum	15	Weightlifti	Middlewe	29228	0	FALSE	Middlewe	1960 Sum	8 Septem	Pa
14	82227	Marlies Rc	Female	22026			East Gern	GDR	Marlies Rc	1980 Wint	41	Cross Cou	5 kilometr	1992	0	FALSE	5 kilometr	1980 Wint	15 Februa	Mt
15	93334	Craig Hutc	Male	27540	198	97	Canada	CAN	0	2000 Sum	25	Swimming	50 metres	8336	0	FALSE	50 metres	2000 Sum	21 â€“ 22 :	Sy
16	146013	Raquel Qu	Female	36589	167	56	Portugal	POR	0	2020 Sum	61	Cycling Mc	Cross-Cou	19001705	0	FALSE	Cross-Cou	2020 Sum	27 July 20:	Izu
17	109912	Vyachesla	Male	31768	170	65	Russian F	RUS	0	2006 Wint	49	Short Trac	500 metre	838	0	FALSE	500 metre	2006 Wint	22 â€“ 25	Pa
18	37019	Philippe Lc	Male	24636	189	85	France	FRA	0	1992 Sum	23	Rowing	Coxed Fou	159262	0	TRUE	Coxed Fou	1992 Sum	27 July â€'	La
19	22885	Rudy Kuge	Male	10451	187	86	Luxembo	LUX	0	1960 Sum	15	Fencing	Épée, Tear	88778	0	TRUE	Ã‰pÃ©e,	1960 Sum	9 Septem	Pa
20	95497	Yoshihiro	Male	11088	173	70	Japan	JPN	0	1960 Wint	36	Ice Hockey	Ice Hockey	20243	0	TRUE	Ice Hockey	1960 Wint	19 â€“ 28	Bly
21	76	Roper Bar	Male	24 November 1873			Great Brit	GBR	Roper Bar	1908 Sum	5	Tennis	Singles, M	44210	0	FALSE	Singles, M	1908 Sum	6 â€“ 11 Ju	All
22	47023	Lorna Frar	Female	7398			Great Brit	GBR	Lorna Frar	1936 Sum	11	Swimming	100 metre	5109	0	FALSE	100 metre	1936 Sum	11 â€“ 13 .	Scl
23	42069	Isabelle H.	Female	21681	167	67	France	FRA	0	1988 Sum	22	Shooting	Small-Bor	51819	0	FALSE	Small-Bor	1988 Sum	21 Septem	Ta
24	39033	Sjoerd Wa	Male	14366	189	76	Netherlar	NED	0	1964 Sum	16	Rowing	Coxless Fc	158436	0	TRUE	Coxless Fc	1964 Sum	11 â€“ 15	To
25	115157	Mariya Ya	Female	29957	174	81	Russian F	RUS	Personal E	2008 Sum	53	Athletics	Javelin Th	257805	0	FALSE	Javelin Th	2008 Sum	19 â€“ 21 .	Be
26	207	VirÃ¡g Csu	Female	26613	172	63	Hungary	HUN	VirÃ¡g Csu	1996 Sum	24	Tennis	Singles, W	45549	0	FALSE	Singles, W	1996 Sum	23 July â€'	Sto
27	99106	Ronny Yea	Male	19222	181	70	United St	USA	Ronny Yea	1972 Wint	39	Cross Cou	15 kilomet	1960	0	FALSE	15 kilomet	1972 Wint	7 Februar	Ma
28	126257	Shane Sm	Male	29858	184	79	New Zeal	NZL	0	2012 Sum	54	Football	Football, M	312000	0	TRUE	Football, M	2012 Sum	26 July â€'	Cit

Note: Some Extra Steps are performed in our project



	athlete_id	name	sex	born	height	weight	country	country_noc	description	edition...10
1	65649	Ivanka Bonova	Female	17992	166	55	Bulgaria	BUL	Personal Best: 400 – 53.54 (1980).	1976 Summer Olympics
2	112510	Nataliya Uryadova	Female	28199	184	70	Russian Federation	RUS	0	2008 Summer Olympics
3	114973	Essa Ismail Rashed	Male	31760	165	55	Qatar	QAT	Personal Best: 10000 – 27:20.97 (2006).	2008 Summer Olympics
4	30359	Péter Boros	Male	2934	NA	NA	Hungary	HUN	Between 1927 and 1938, Péter Boros competed as a gymna...	1932 Summer Olympics
5	50557	Rudolf Piowatý	Male	119	NA	NA	Czechoslovakia	TCH	Rudolf Piowaty joined the Czechoslovak military near the en...	1924 Summer Olympics
6	146111	Svetlana Kholomina	Female	35743	NA	NA	ROC	ROC	0	2020 Summer Olympics
7	133041	Vincent Riendeau	Male	35412	178	68	Canada	CAN	0	2016 Summer Olympics
8	110425	Tanja Morel	Female	27671	164	58	Switzerland	SUI	0	2006 Winter Olympics
9	110705	Maksim Shabalin	Male	29976	183	76	Russian Federation	RUS	0	2006 Winter Olympics
10	54541	Gé Regter	Male	5910	NA	NA	Netherlands	NED	0	1936 Summer Olympics
11	22721	Aristide Pontenani	Male	NA	NA	NA	Italy	ITA	0	1912 Summer Olympics
12	56266	Go Yeong-Chang	Male	9577	167	75	Republic of Korea	KOR	0	1960 Summer Olympics
13	82227	Marlies Rostock	Female	22026	NA	NA	East Germany	GDR	Marlies Rostock won gold in the cross-country relay at Lake ...	1980 Winter Olympics
14	93334	Craig Hutchison	Male	27540	198	97	Canada	CAN	0	2000 Summer Olympics
15	146013	Raquel Queirós	Female	36589	167	56	Portugal	POR	0	2020 Summer Olympics
16	109912	Vyacheslav Kurginyan	Male	31768	170	65	Russian Federation	RUS	0	2006 Winter Olympics
17	37019	Philippe Lot	Male	24636	189	85	France	FRA	0	1992 Summer Olympics
18	22885	Rudy Kugeler	Male	10451	187	86	Luxembourg	LUX	0	1960 Summer Olympics
19	95497	Yoshihiro Miyazaki	Male	11088	173	70	Japan	JPN	0	1960 Winter Olympics

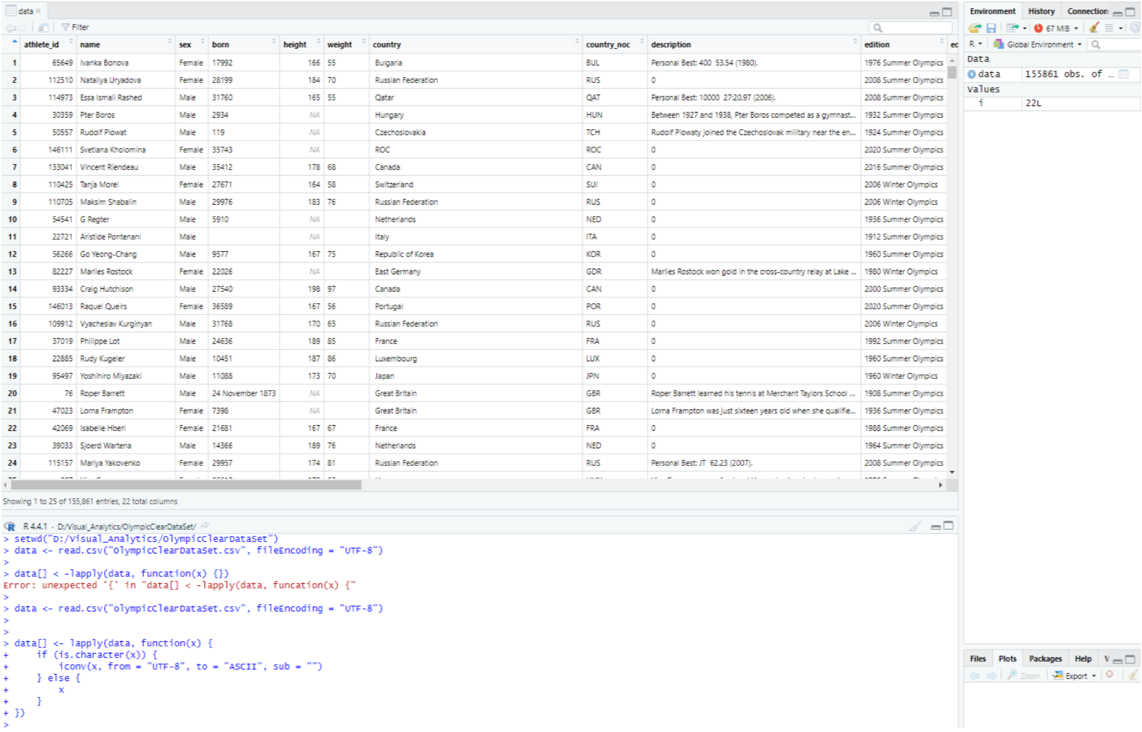
Showing 1 to 20 of 155,861 entries, 22 total columns

```
Data
data      155861 obs. of ...
Values
  i       22L
```

```
R 4.4.1 · D:/Visual_Analytics/RFinalProject/OlympicClearDataSet/
> setwd("D:/Visual_Analytics/RFinalProject/OlympicClearDataSet")
> data <- read.csv('OlympicClearDataSet.csv')
Error in type.convert.default(data[[i]], as.is = as.is[i], dec = dec,  :
  invalid multibyte string at '<d7> 40'
> data <- read.csv("OlympicClearDataSet", fileEncoding = "UTF-8")
Error in file(file, "rt", encoding = fileEncoding) :
  cannot open the connection
In addition: Warning message:
In file(file, "rt", encoding = fileEncoding) :
  cannot open file 'OlympicClearDataSet': No such file or directory
> data <- read.csv("OlympicClearDataSet", fileEncoding = "windows-1252")
Error in file(file, "rt", encoding = fileEncoding) :
  cannot open the connection
In addition: Warning message:
In file(file, "rt", encoding = fileEncoding) :
  cannot open file 'OlympicClearDataSet': No such file or directory
> grep("[^\x20-\x7E]", data$column_name, value = TRUE)
character(0)
>
> data$column_name <- iconv(data$column_name, from = "UTF-8", to = "ASCII", sub = "")
Error in `$<-.data.frame`(`*tmp*`, column_name, value = character(0)) :
  replacement has 0 rows, data has 155861
> library(readr)
> data <- read_csv("OlympicClearDataSet", locale = locale(encoding = "UTF-8"))
Error: 'OlympicClearDataSet' does not exist in current working directory ('D:/Visual_Analytics/RFinalProject/OlympicClearDataSet').
> library(readr)
> data <- read_csv("OlympicClearDataSet.csv", locale = locale(encoding = "UTF-8"))
New names:
• `edition` -> `edition...10`
```

```
> setwd("D:/Visual_Analytics/RFinalProject/OlympicClearDataSet")
> data <- read.csv('OlympicClearDataSet.csv')
Error in type.convert.default(data[[i]], as.is = as.is[i], dec = dec,  :
  invalid multibyte string at '<d7> 40'
> data <- read.csv("OlympicClearDataSet", fileEncoding = "UTF-8")
Error in file(file, "rt", encoding = fileEncoding) :
  cannot open the connection
In addition: Warning message:
In file(file, "rt", encoding = fileEncoding) :
  cannot open file 'OlympicClearDataSet': No such file or directory
> data <- read.csv("OlympicClearDataSet", fileEncoding = "windows-1252")
Error in file(file, "rt", encoding = fileEncoding) :
  cannot open the connection
In addition: Warning message:
In file(file, "rt", encoding = fileEncoding) :
  cannot open file 'OlympicClearDataSet': No such file or directory
> grep("[^\x20-\x7E]", data$column_name, value = TRUE)
character(0)
>
> data$column_name <- iconv(data$column_name, from = "UTF-8", to = "ASCII", sub = "")
Error in `$<-.data.frame`(`*tmp*`, column_name, value = character(0)) :
  replacement has 0 rows, data has 155861
> library(readr)
> data <- read_csv("OlympicClearDataSet", locale = locale(encoding = "UTF-8"))
Error: 'OlympicClearDataSet' does not exist in current working directory ('D:/Visual_Analytics/RFinalProject/OlympicClearDataSet').
> library(readr)
> data <- read_csv("OlympicClearDataSet.csv", locale = locale(encoding = "UTF-8"))
New names:
```

Once the file is saved into CSV-UTF, Again I uploaded my CSV file to R Studio



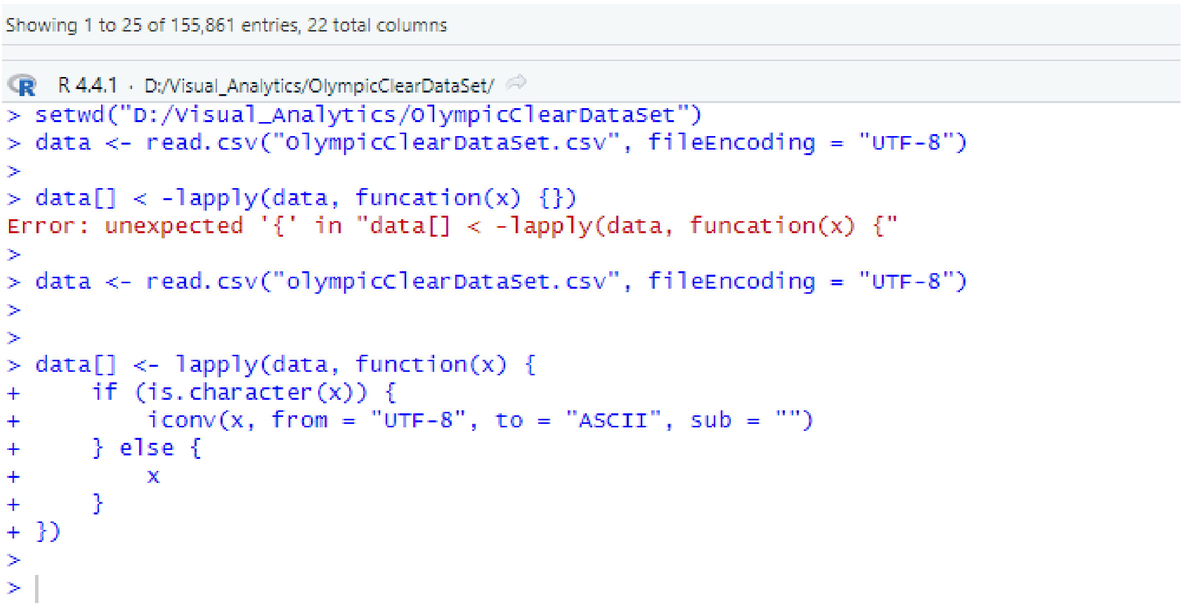
	athlete_id	name	sex	born	height	weight	country	country_noc	description	edition	ec
1	65649	Ivanka Bonova	Female	17992	166	55	Bulgaria	BUL	Personal Best: 400 53.54 (1980).	1976 Summer Olympics	
2	112510	Nataliya Uryadova	Female	28199	184	70	Russian Federation	RUS	0	2008 Summer Olympics	
3	114973	Essa Ismail Rashed	Male	31760	165	55	Qatar	QAT	Personal Best: 10000 27:20.97 (2006).	2008 Summer Olympics	
4	30359	Pter Boros	Male	2934	NA		Hungary	HUN	Between 1927 and 1938, Pter Boros competed as a gymnast...	1932 Summer Olympics	
5	50557	Rudolf Piowat	Male	119	NA		Czechoslovakia	TCH	Rudolf Piowaty joined the Czechoslovak military near the en...	1924 Summer Olympics	
6	146111	Svetlana Kholomina	Female	35743	NA		ROC	ROC	0	2020 Summer Olympics	
7	133041	Vincent Riendeau	Male	35412	178	68	Canada	CAN	0	2016 Summer Olympics	
8	110425	Tanja Morel	Female	27671	164	58	Switzerland	SUI	0	2006 Winter Olympics	
9	110705	Maksim Shabalin	Male	29976	183	76	Russian Federation	RUS	0	2006 Winter Olympics	
10	54541	G Regter	Male	5910	NA		Netherlands	NED	0	1936 Summer Olympics	
11	22721	Aristide Pontenani	Male		NA		Italy	ITA	0	1912 Summer Olympics	
12	56266	Go Yeong-Chang	Male	9577	167	75	Republic of Korea	KOR	0	1960 Summer Olympics	
13	82227	Marlies Rostock	Female	22026	NA		East Germany	GDR	Marlies Rostock won gold in the cross-country relay at Lake ...	1980 Winter Olympics	
14	93334	Craig Hutchison	Male	27540	198	97	Canada	CAN	0	2000 Summer Olympics	
15	146013	Raquel Queirs	Female	36589	167	56	Portugal	POR	0	2020 Summer Olympics	
16	109912	Vyacheslav Kurginyan	Male	31768	170	65	Russian Federation	RUS	0	2006 Winter Olympics	
17	37019	Philippe Lot	Male	24636	189	85	France	FRA	0	1992 Summer Olympics	
18	22885	Rudy Kugeler	Male	10451	187	86	Luxembourg	LUX	0	1960 Summer Olympics	
19	95497	Yoshihiro Miyazaki	Male	11088	173	70	Japan	JPN	0	1960 Winter Olympics	
20	76	Roper Barrett	Male	24 November 1873	NA		Great Britain	GBR	Roper Barrett learned his tennis at Merchant Taylors School ...	1908 Summer Olympics	
21	47023	Lorna Frampton	Female	7398	NA		Great Britain	GBR	Lorna Frampton was just sixteen years old when she qualifie...	1936 Summer Olympics	
22	42069	Isabelle Hberl	Female	21681	167	67	France	FRA	0	1988 Summer Olympics	
23	39033	Sjoerd Wartena	Male	14366	189	76	Netherlands	NED	0	1964 Summer Olympics	
24	115157	Mariya Yakovenko	Female	29957	174	81	Russian Federation	RUS	Personal Best: JT 62.23 (2007).	2008 Summer Olympics	

Showing 1 to 25 of 155,861 entries, 22 total columns

```
Data
data      155861 obs. of …
Values
  i       22L
```

```
R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/
> setwd("D:/Visual_Analytics/OlympicClearDataSet")
> data <- read.csv("OlympicClearDataSet.csv", fileEncoding = "UTF-8")
>
> data[] < -lapply(data, funcation(x) {})
Error: unexpected '{' in "data[] < -lapply(data, funcation(x) {"
>
> data <- read.csv("olympicClearDataSet.csv", fileEncoding = "UTF-8")
>
>
> data[] <- lapply(data, function(x) {
+     if (is.character(x)) {
+         iconv(x, from = "UTF-8", to = "ASCII", sub = "")
+     } else {
+         x
+     }
+ })
>
```

Trying to read the dataset and convert it from UTF to ASCII


Showing 1 to 25 of 155,861 entries, 22 total columns

```
R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/
> setwd("D:/Visual_Analytics/OlympicClearDataSet")
> data <- read.csv("OlympicClearDataSet.csv", fileEncoding = "UTF-8")
>
> data[] < -lapply(data, funcation(x) {})
Error: unexpected '{' in "data[] < -lapply(data, funcation(x) {"
>
> data <- read.csv("olympicClearDataSet.csv", fileEncoding = "UTF-8")
>
>
> data[] <- lapply(data, function(x) {
+     if (is.character(x)) {
+         iconv(x, from = "UTF-8", to = "ASCII", sub = "")
+     } else {
+         x
+     }
+ })
>
> 
```

Data Cleaning:

Steps Followed in the below Screen shots:

Sessions > SetworkingDirectory > Choose Directory

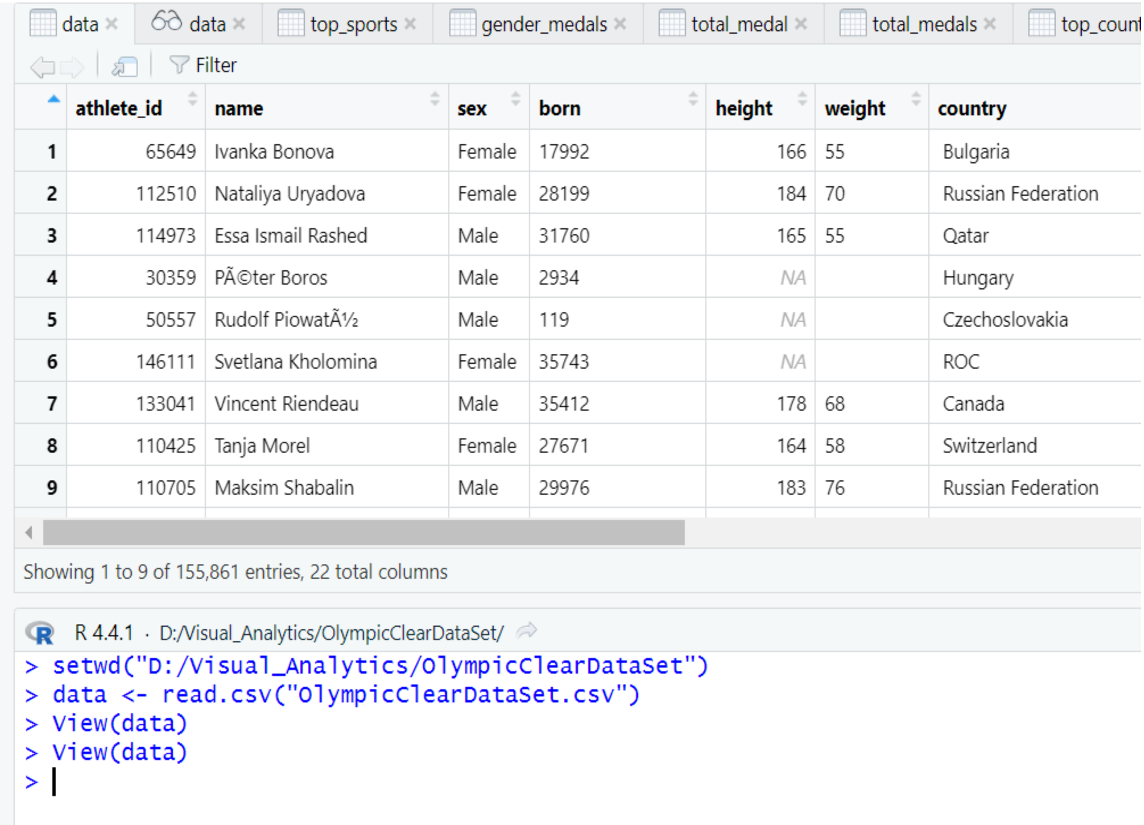
In my case I have saved the CSV file under

D:/Visual_Anlytics/RFinalProject/OlympicClearDataSet

OlympicClearDataSet is loaded into R Studio:

Before Data Cleaning :

Before data cleaning, the dataset is simply loaded into R Studio. In the initial stage, I review the dataset by reading and viewing it. As shown in the screenshot above, there is a significant amount of missing data, duplicate data, and inconsistent data

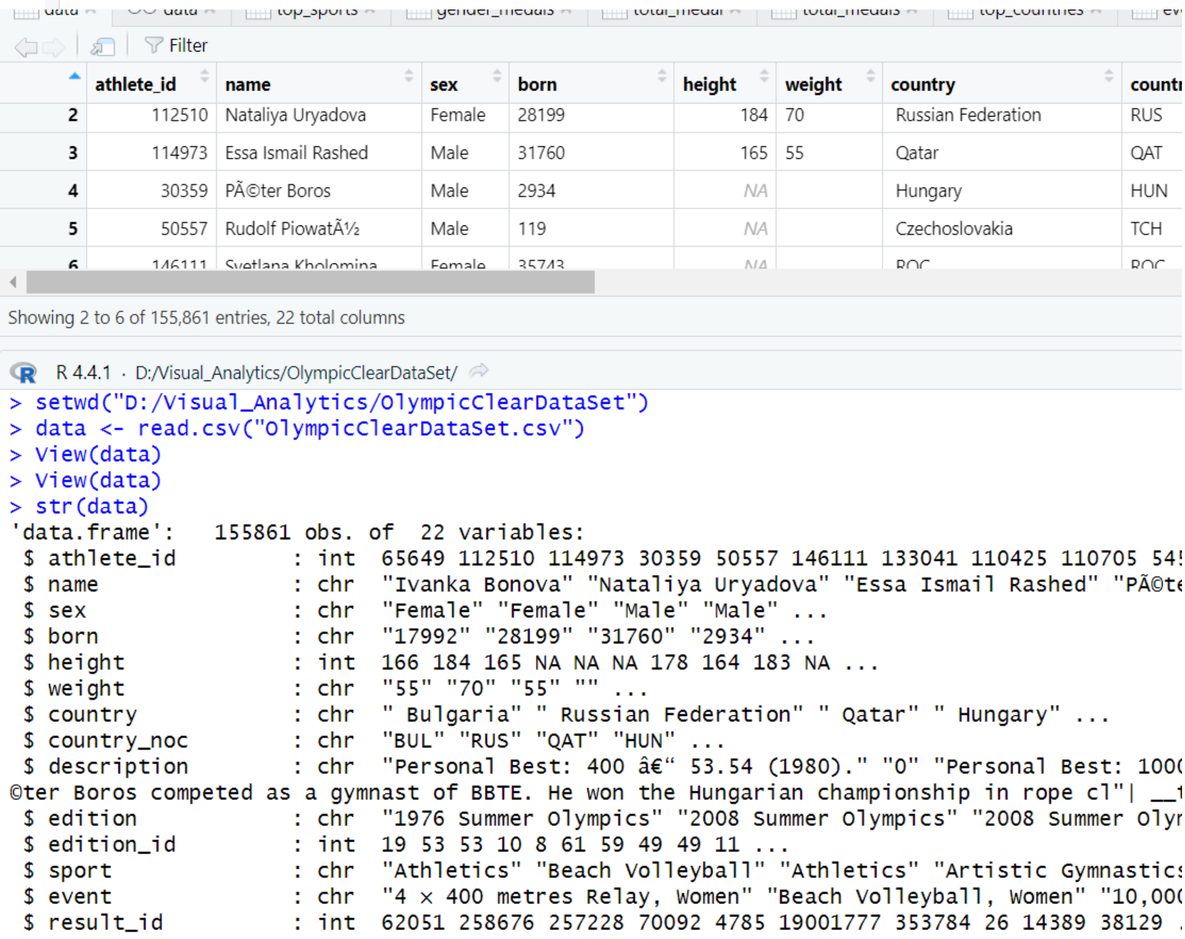


	athlete_id	name	sex	born	height	weight	country
1	65649	Ivanka Bonova	Female	17992	166	55	Bulgaria
2	112510	Nataliya Uryadova	Female	28199	184	70	Russian Federation
3	114973	Essa Ismail Rashed	Male	31760	165	55	Qatar
4	30359	PÃ©ter Boros	Male	2934	NA		Hungary
5	50557	Rudolf PiowatÃ½	Male	119	NA		Czechoslovakia
6	146111	Svetlana Kholomina	Female	35743	NA		ROC
7	133041	Vincent Riendeau	Male	35412	178	68	Canada
8	110425	Tanja Morel	Female	27671	164	58	Switzerland
9	110705	Maksim Shabalin	Male	29976	183	76	Russian Federation

Showing 1 to 9 of 155,861 entries, 22 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> setwd("D:/Visual_Analytics/OlympicClearDataSet")
> data <- read.csv("OlympicClearDataSet.csv")
> View(data)
> View(data)
> 
```



	athlete_id	name	sex	born	height	weight	country	countr
2	112510	Nataliya Uryadova	Female	28199	184	70	Russian Federation	RUS
3	114973	Essa Ismail Rashed	Male	31760	165	55	Qatar	QAT
4	30359	PÃ©ter Boros	Male	2934	NA		Hungary	HUN
5	50557	Rudolf PiowatÃ½	Male	119	NA		Czechoslovakia	TCH
6	146111	Svetlana Kholomina	Female	35743	NA		ROC	ROC

Showing 2 to 6 of 155,861 entries, 22 total columns

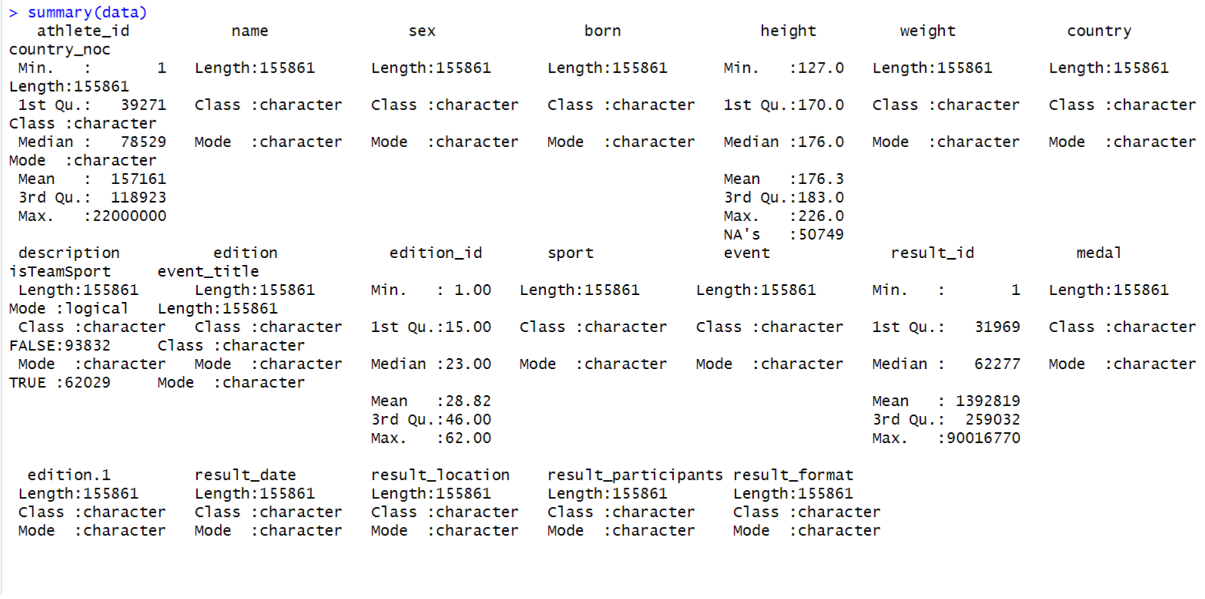
R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> setwd("D:/Visual_Analytics/OlympicClearDataSet")
> data <- read.csv("OlympicClearDataSet.csv")
> View(data)
> View(data)
> str(data)
'data.frame':   155861 obs. of  22 variables:
 $ athlete_id         : int  65649 112510 114973 30359 50557 146111 133041 110425 110705 54!
 $ name               : chr  "Ivanka Bonova" "Nataliya Uryadova" "Essa Ismail Rashed" "PÃ©te
 $ sex                : chr  "Female" "Female" "Male" "Male" ...
 $ born               : chr  "17992" "28199" "31760" "2934" ...
 $ height             : int  166 184 165 NA NA NA 178 164 183 NA ...
 $ weight             : chr  "55" "70" "55" "" ...
 $ country            : chr  " Bulgaria" " Russian Federation" " Qatar" " Hungary" ...
 $ country_noc        : chr  "BUL" "RUS" "QAT" "HUN" ...
 $ description        : chr  "Personal Best: 400 â€“ 53.54 (1980)." "0" "Personal Best: 100(
©ter Boros competed as a gymnast of BBTE. He won the Hungarian championship in rope cl"| __t
 $ edition            : chr  "1976 Summer Olympics" "2008 Summer Olympics" "2008 Summer Olyr
 $ edition_id         : int  19 53 53 10 8 61 59 49 49 11 ...
 $ sport              : chr  "Athletics" "Beach Volleyball" "Athletics" "Artistic Gymnastics
 $ event              : chr  "4 × 400 metres Relay, Women" "Beach Volleyball, Women" "10,00(
 $ result_id          : int  62051 258676 257228 70092 4785 19001777 353784 26 14389 38129 .
```

Code

```
Data <- read.csv("OlympicClearDataset.csv")
Str(data)
Summary(data)
```

This linereads aCSV (Comma-Separated Values) file named OlympicClearDataset.csv into R and stores it in a variable named Data. A function in R used to import data from a CSV file into a dataframe. It automatically converts the data into tabular format where rows represent observations and columns represent variables. This is the name of the dataframe where the dataset is stored.



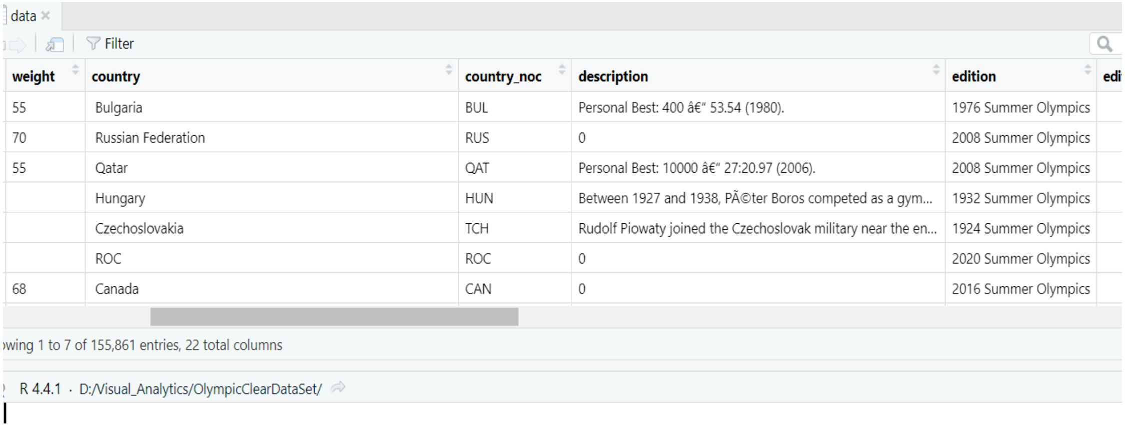
```
> summary(data)
   athlete_id          name                sex                born               height          weight             country
country_noc
 Min.   :       1   Length:155861      Length:155861      Length:155861      Min.   :127.0   Length:155861      Length:155861
Length:155861
 1st Qu.:   39271   Class :character   Class :character   Class :character   1st Qu.:170.0   Class :character   Class :character
Class :character
 Median :   78529   Mode  :character   Mode  :character   Mode  :character   Median :176.0   Mode  :character   Mode  :character
Mode  :character
 Mean   :  157161                                                            Mean   :176.3
 3rd Qu.:  118923                                                            3rd Qu.:183.0
 Max.   :22000000                                                            Max.   :226.0
                                                                             NA's   :50749
 description          edition            edition_id       sport              event              result_id           medal
isTeamSport     event_title
 Length:155861      Length:155861      Min.   : 1.00   Length:155861      Length:155861      Min.   :       1   Length:155861
Mode :logical   Length:155861
 Class :character   Class :character   1st Qu.:15.00   Class :character   Class :character   1st Qu.:   31969   Class :character
FALSE:93832     Class :character
 Mode  :character   Mode  :character   Median :23.00   Mode  :character   Mode  :character   Median :   62277   Mode  :character
TRUE :62029     Mode  :character
                                       Mean   :28.82                                         Mean   : 1392819
                                       3rd Qu.:46.00                                         3rd Qu.:  259032
                                       Max.   :62.00                                         Max.   :90016770

  edition.1         result_date        result_location    result_participants result_format
 Length:155861      Length:155861      Length:155861      Length:155861       Length:155861
 Class :character   Class :character   Class :character   Class :character    Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character    Mode  :character
```

1)Remove description and Special_notes as it was irrelevant

Code

```
Data <- data [, !(names(data) %in% c("description", "special_notes")]]
```

Before



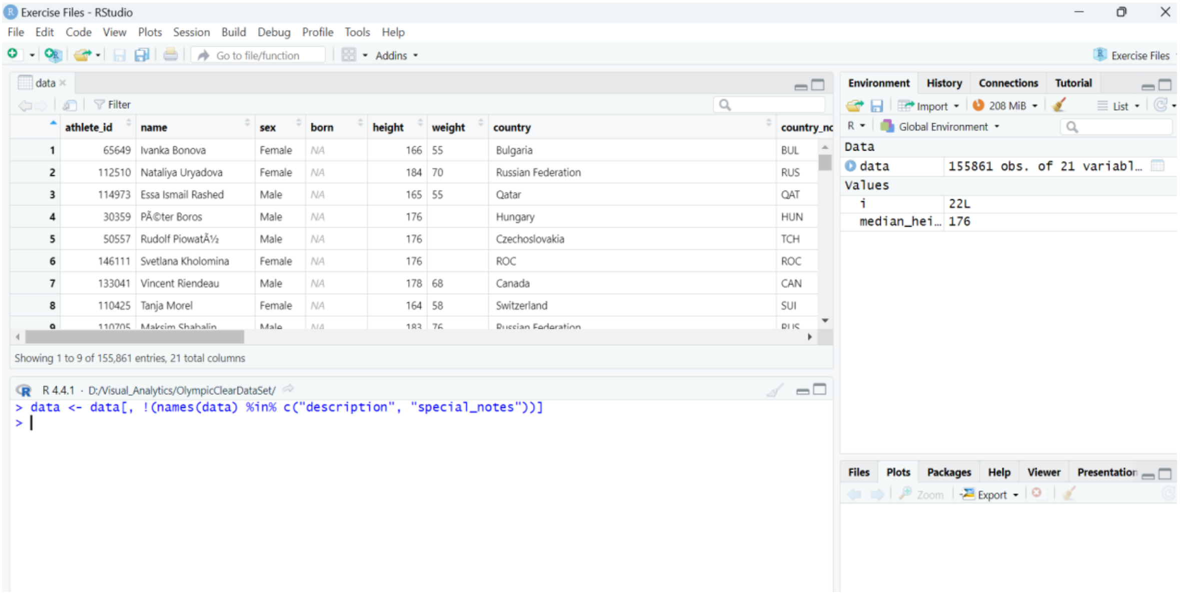
weight	country	country_noc	description	edition	edi
55	Bulgaria	BUL	Personal Best: 400 â€" 53.54 (1980).	1976 Summer Olympics	
70	Russian Federation	RUS	0	2008 Summer Olympics	
55	Qatar	QAT	Personal Best: 10000 â€" 27:20.97 (2006).	2008 Summer Olympics	
	Hungary	HUN	Between 1927 and 1938, PÃ©ter Boros competed as a gym...	1932 Summer Olympics	
	Czechoslovakia	TCH	Rudolf Piowaty joined the Czechoslovak military near the en...	1924 Summer Olympics	
	ROC	ROC	0	2020 Summer Olympics	
68	Canada	CAN	0	2016 Summer Olympics	

ıwing 1 to 7 of 155,861 entries, 22 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

A

After



Exercise Files - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Exercise Files
data
Filter
Environment History Connections Tutorial
Import
208 MiB
List
R
Global Environment
Files Plots Packages Help Viewer Presentation
Zoom
Export

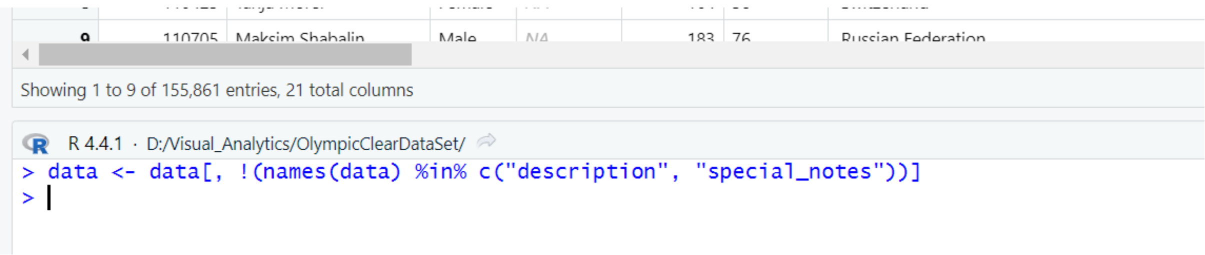
	athlete_id	name	sex	born	height	weight	country	country_nc
1	65649	Ivanka Bonova	Female	NA	166	55	Bulgaria	BUL
2	112510	Nataliya Uryadova	Female	NA	184	70	Russian Federation	RUS
3	114973	Essa Ismail Rashed	Male	NA	165	55	Qatar	QAT
4	30359	PÃ©ter Boros	Male	NA	176		Hungary	HUN
5	50557	Rudolf PiowatÃ½	Male	NA	176		Czechoslovakia	TCH
6	146111	Svetlana Kholomina	Female	NA	176		ROC	ROC
7	133041	Vincent Riendeau	Male	NA	178	68	Canada	CAN
8	110425	Tanja Morel	Female	NA	164	58	Switzerland	SUI
9	110705	Maksim Shabalin	Male	NA	183	76	Russian Federation	RUS

Showing 1 to 9 of 155,861 entries, 21 total columns

Data	
data	155861 obs. of 21 variabl…
Values	
i	22L
median_hei…	176

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> data <- data[, !(names(data) %in% c("description", "special_notes"))]
> |
```



9	110705	Maksim Shabalin	Male	NA	183	76	Russian Federation

Showing 1 to 9 of 155,861 entries, 21 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

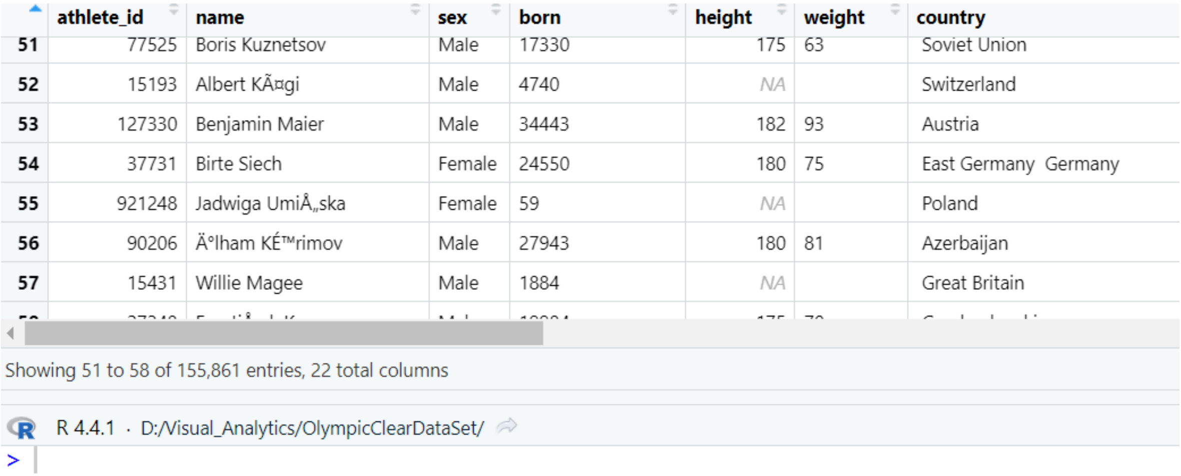
```
> data <- data[, !(names(data) %in% c("description", "special_notes"))]
> |
```

Explanation

This function retrieves the column names of the dataframe data. This operator checks if each column name is in the vector c("description", "special_notes"). It returns TRUE for column names that match and FALSE otherwise. The negation operator ! reverses the logical values Columns that match description or special_notes will have TRUE, and applying ! turns them into FALSE. Columns that do not match remain TRUE. The square brackets [,] are used to subset the dataframe. The first argument (before the comma) specifies rows (here it is blank, meaning all rows).The second argument (after the comma) specifies columns to keep. The code ensures that only columns **not** in c("description"special_notes") are retained.

2)Remove Duplicates

Before

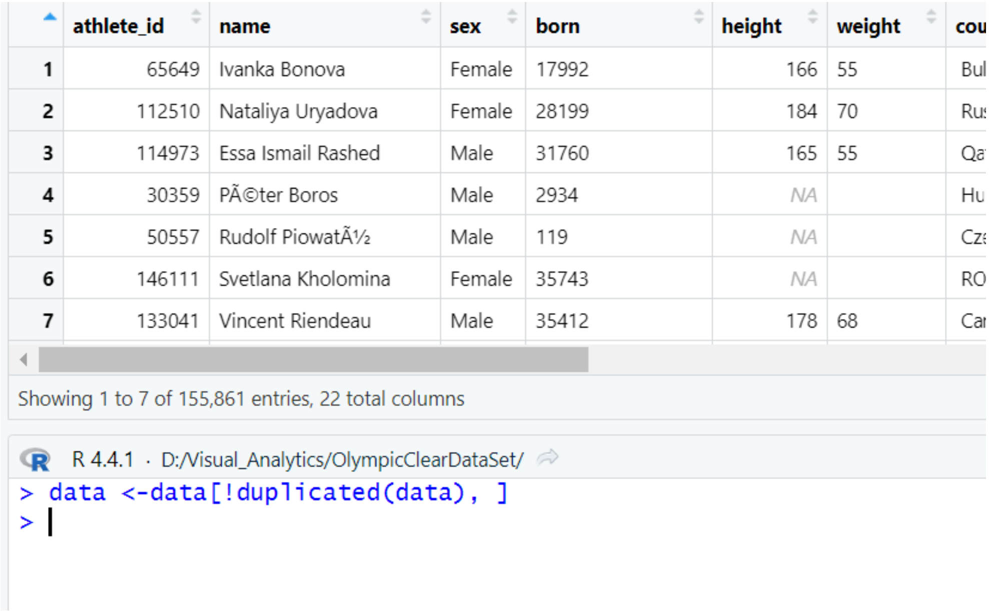


	athlete_id	name	sex	born	height	weight	country
51	77525	Boris Kuznetsov	Male	17330	175	63	Soviet Union
52	15193	Albert KÃ¤gi	Male	4740	NA		Switzerland
53	127330	Benjamin Maier	Male	34443	182	93	Austria
54	37731	Birte Siech	Female	24550	180	75	East Germany Germany
55	921248	Jadwiga UmiÅ„ska	Female	59	NA		Poland
56	90206	Ä°lham KÉ™rimov	Male	27943	180	81	Azerbaijan
57	15431	Willie Magee	Male	1884	NA		Great Britain

Showing 51 to 58 of 155,861 entries, 22 total columns

```
R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/
>
```

After



	athlete_id	name	sex	born	height	weight	cou
1	65649	Ivanka Bonova	Female	17992	166	55	Bul
2	112510	Nataliya Uryadova	Female	28199	184	70	Ru
3	114973	Essa Ismail Rashed	Male	31760	165	55	Qa
4	30359	PÃ©ter Boros	Male	2934	NA		Hu
5	50557	Rudolf PiowatÃ½	Male	119	NA		Cz
6	146111	Svetlana Kholomina	Female	35743	NA		RO
7	133041	Vincent Riendeau	Male	35412	178	68	Ca

Showing 1 to 7 of 155,861 entries, 22 total columns

```
R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/
> data <-data[!duplicated(data), ]
>
```

```
Data <- data[!duplicated(data),    ]
```

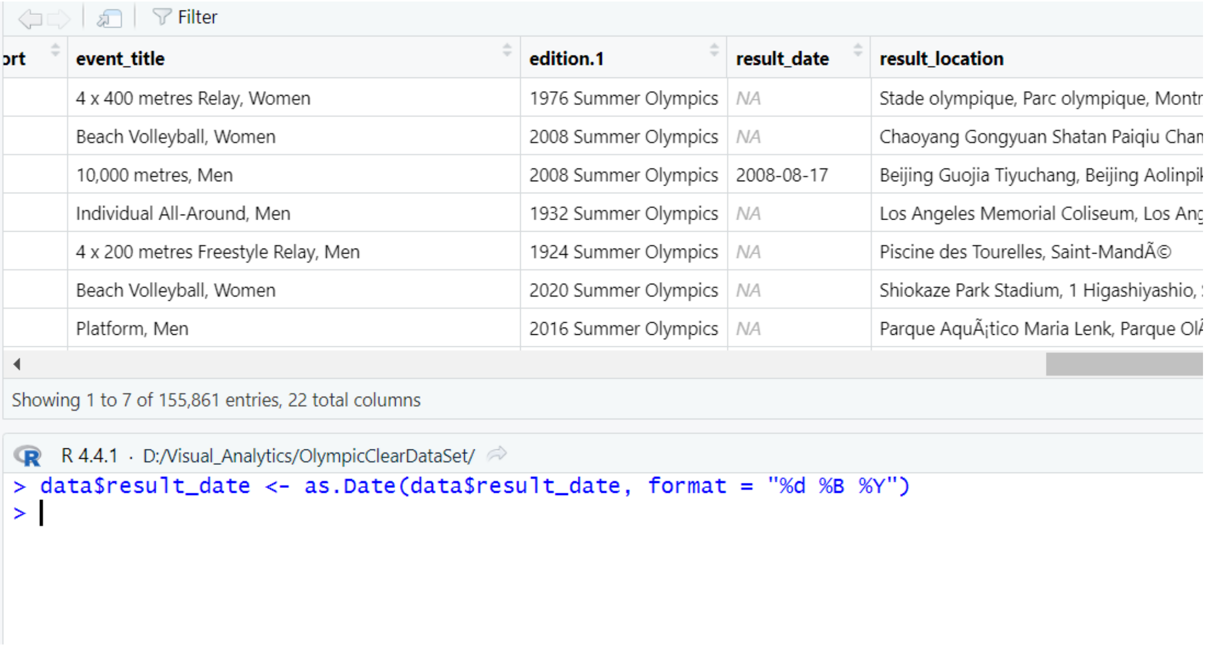
Explanation

This function identifies duplicate rows in the dataframe data. It returns a logical vector where, TRUE indicates a duplicate row (i.e., a row with the same values as a previous row). FALSE indicates a unique row (i.e., no previous row has identical values).

The negation operator ! reverses the logical values TRUE becomes FALSE. FALSE becomes TRUE As a result, this keeps only the first occurrence of duplicate rows while marking subsequent duplicates. This subsets the dataframe data to include only the rows where !duplicated(data) is TRUE. The square brackets [,] are used for subsetting:The condition !duplicated(data) applies to the rows. The empty column argument after the comma (,) indicates all columns are kept.

3)Result Formatting

Before



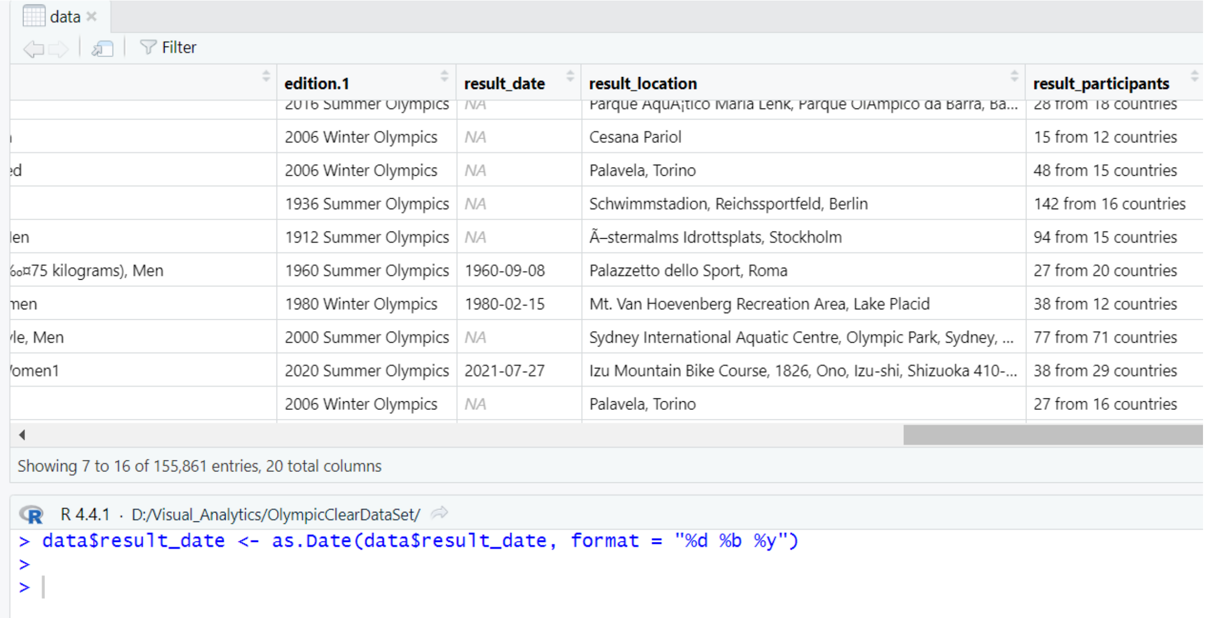
ort	event_title	edition.1	result_date	result_location
	4 x 400 metres Relay, Women	1976 Summer Olympics	NA	Stade olympique, Parc olympique, Montr
	Beach Volleyball, Women	2008 Summer Olympics	NA	Chaoyang Gongyuan Shatan Paiqiu Chan
	10,000 metres, Men	2008 Summer Olympics	2008-08-17	Beijing Guojia Tiyuchang, Beijing Aolinpik
	Individual All-Around, Men	1932 Summer Olympics	NA	Los Angeles Memorial Coliseum, Los Ang
	4 x 200 metres Freestyle Relay, Men	1924 Summer Olympics	NA	Piscine des Tourelles, Saint-MandÃ©
	Beach Volleyball, Women	2020 Summer Olympics	NA	Shiokaze Park Stadium, 1 Higashiyashio,
	Platform, Men	2016 Summer Olympics	NA	Parque AquÃ¡tico Maria Lenk, Parque Olí

Showing 1 to 7 of 155,861 entries, 22 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> data$result_date <- as.Date(data$result_date, format = "%d %B %Y")
> 
```

After



	edition.1	result_date	result_location	result_participants
	2016 Summer Olympics	NA	Parque AquÃ¡tico Maria Lenk, Parque OlÃmpico da Barra, Ba...	28 from 18 countries
ı	2006 Winter Olympics	NA	Cesana Pariol	15 from 12 countries
ed	2006 Winter Olympics	NA	Palavela, Torino	48 from 15 countries
	1936 Summer Olympics	NA	Schwimmstadion, Reichssportfeld, Berlin	142 from 16 countries
len	1912 Summer Olympics	NA	Ã–stermalms Idrottsplats, Stockholm	94 from 15 countries
‰¤75 kilograms), Men	1960 Summer Olympics	1960-09-08	Palazzetto dello Sport, Roma	27 from 20 countries
men	1980 Winter Olympics	1980-02-15	Mt. Van Hoevenberg Recreation Area, Lake Placid	38 from 12 countries
le, Men	2000 Summer Olympics	NA	Sydney International Aquatic Centre, Olympic Park, Sydney, ...	77 from 71 countries
omen1	2020 Summer Olympics	2021-07-27	Izu Mountain Bike Course, 1826, Ono, Izu-shi, Shizuoka 410-...	38 from 29 countries
	2006 Winter Olympics	NA	Palavela, Torino	27 from 16 countries

Showing 7 to 16 of 155,861 entries, 20 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> data$result_date <- as.Date(data$result_date, format = "%d %b %y")
>
> |
```

Code:

```
data  $result_date  <-  as.Date(data$result_date,format  = “%d %b %y”)
```

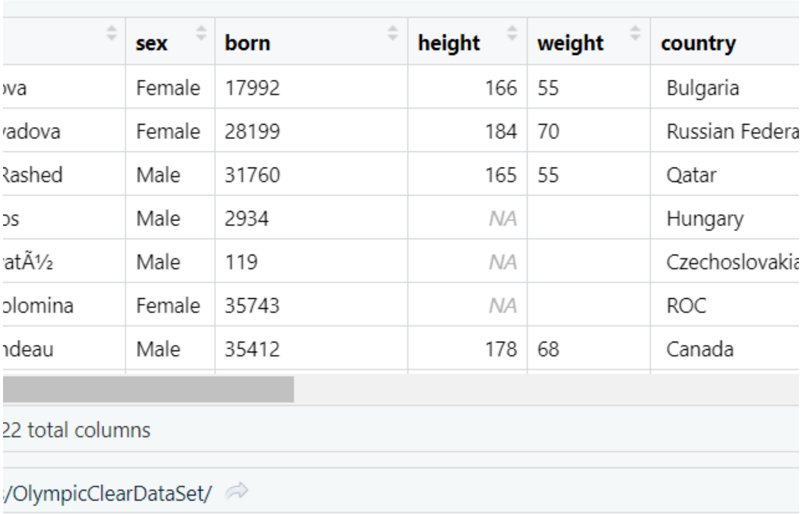
Explanation

We are trying to convert a column named result_date in a dataframe data to a Date object in R using the as.Date Converts a character vector to a Date object.

format = "%d %b %y": %d Day as a number (01-31),%b Abbreviated month name (Jan, Feb),%y Year as a two-digit number.

4) Removing the Athlet born date as it was irrelavent

Before

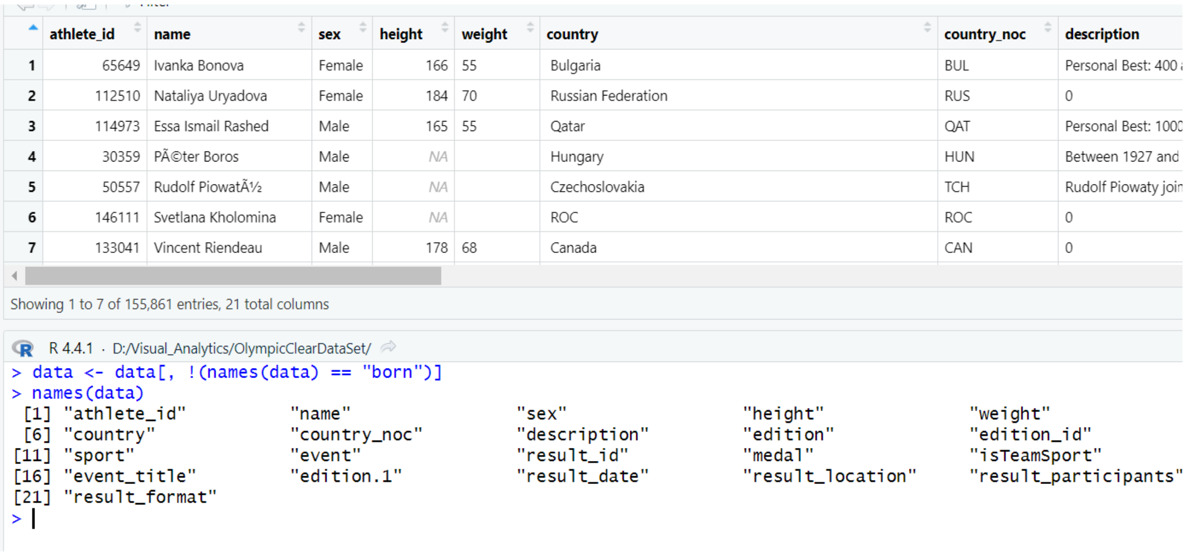


	sex	born	height	weight	country
va	Female	17992	166	55	Bulgaria
adova	Female	28199	184	70	Russian Federa
Rashed	Male	31760	165	55	Qatar
os	Male	2934	NA		Hungary
atÃ½	Male	119	NA		Czechoslovaki
olomina	Female	35743	NA		ROC
ndeau	Male	35412	178	68	Canada

22 total columns

/OlympicClearDataSet/

After



	athlete_id	name	sex	height	weight	country	country_noc	description
1	65649	Ivanka Bonova	Female	166	55	Bulgaria	BUL	Personal Best: 400
2	112510	Nataliya Uryadova	Female	184	70	Russian Federation	RUS	0
3	114973	Essa Ismail Rashed	Male	165	55	Qatar	QAT	Personal Best: 100
4	30359	PÃ©ter Boros	Male	NA		Hungary	HUN	Between 1927 and
5	50557	Rudolf PiowatÃ½	Male	NA		Czechoslovakia	TCH	Rudolf Piowaty joi
6	146111	Svetlana Kholomina	Female	NA		ROC	ROC	0
7	133041	Vincent Riendeau	Male	178	68	Canada	CAN	0

Showing 1 to 7 of 155,861 entries, 21 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> data <- data[, !(names(data) == "born")]
> names(data)
 [1] "athlete_id"          "name"                "sex"                 "height"              "weight"
 [6] "country"             "country_noc"         "description"         "edition"             "edition_id"
[11] "sport"               "event"               "result_id"           "medal"               "isTeamSport"
[16] "event_title"         "edition.1"           "result_date"         "result_location"     "result_participants"
[21] "result_format"
> |
```

Code

```
Write.csv(data,“Dataset_without_Born.csv”, row.name  = FALSE)
```

Explanation:

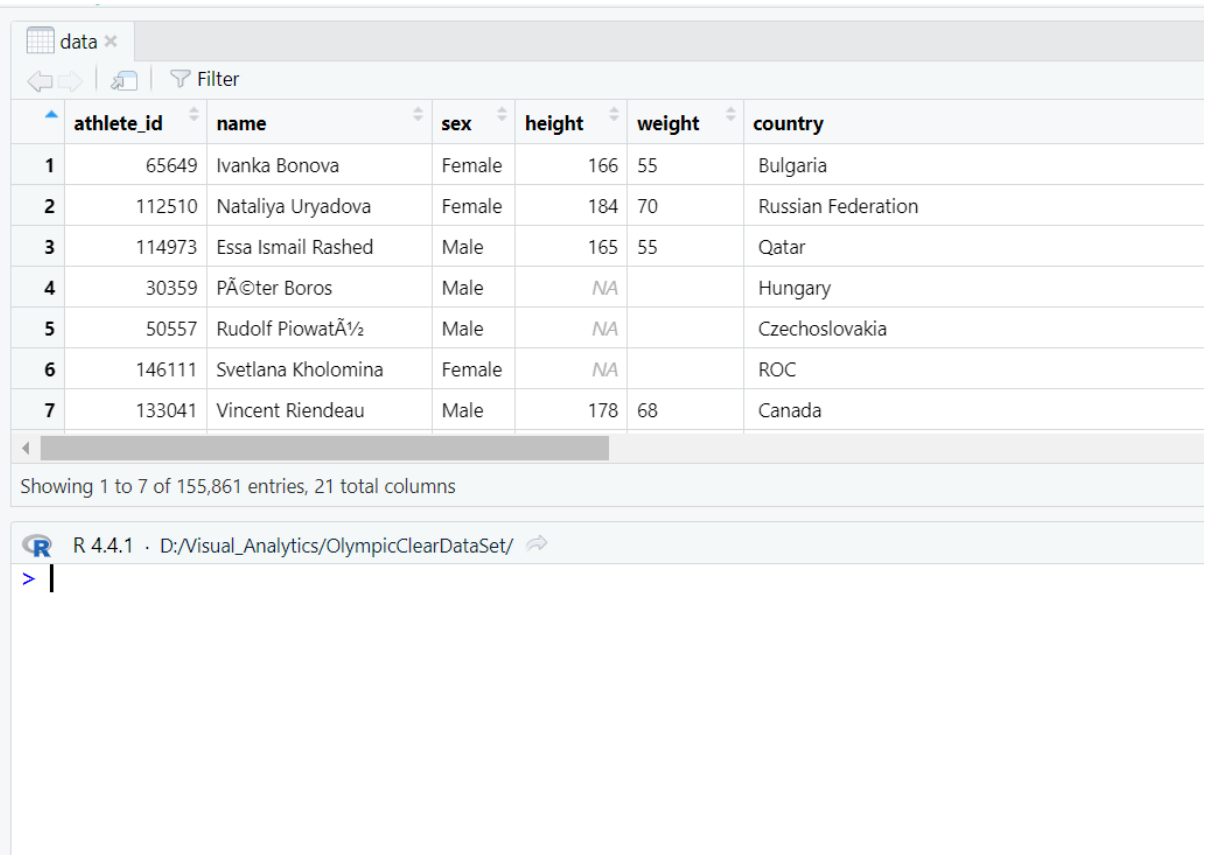
The write.csv() function in R is used to export data frames or matrices to a CSV files are a common format for data exchange. Data: This is the frame you want to explore.

Dataset_without_Born.csv: This is the name of the csv file that will be created. If no path is specified, the file will be saved in the current working directory, which can be checked using getwd()

row.name=FALSE:By default, R includes row names (the first column of the data frame, often indices) when writing a CSV file. Setting row.names = FALSE excludes these row names from the output. If set to TRUE, an additional column would be added to the CSV containing the row names, which might not be desired unless explicitly needed

5) Replace missing Height & Weight

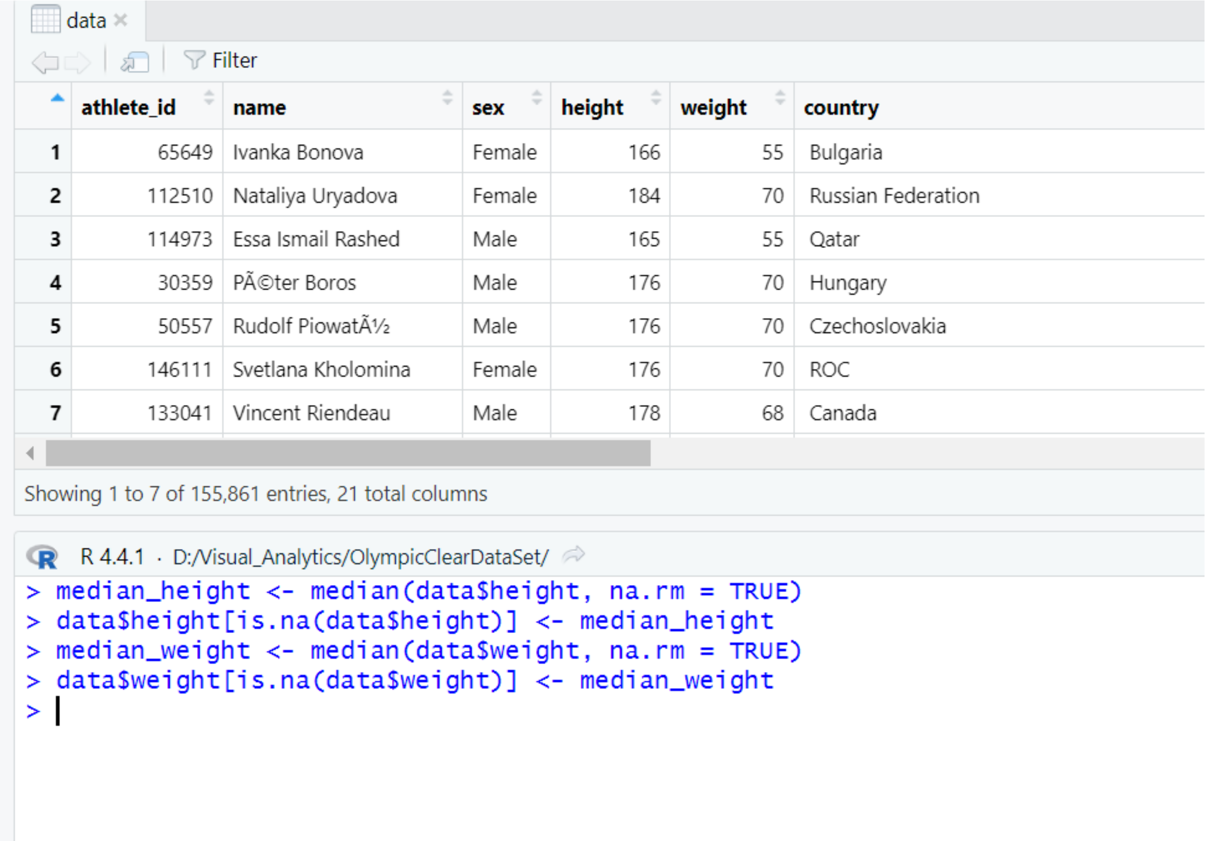
Before:



	athlete_id	name	sex	height	weight	country
1	65649	Ivanka Bonova	Female	166	55	Bulgaria
2	112510	Nataliya Uryadova	Female	184	70	Russian Federation
3	114973	Essa Ismail Rashed	Male	165	55	Qatar
4	30359	PÃ©ter Boros	Male	NA		Hungary
5	50557	Rudolf PiowatÃ½	Male	NA		Czechoslovakia
6	146111	Svetlana Kholomina	Female	NA		ROC
7	133041	Vincent Riendeau	Male	178	68	Canada

Showing 1 to 7 of 155,861 entries, 21 total columns

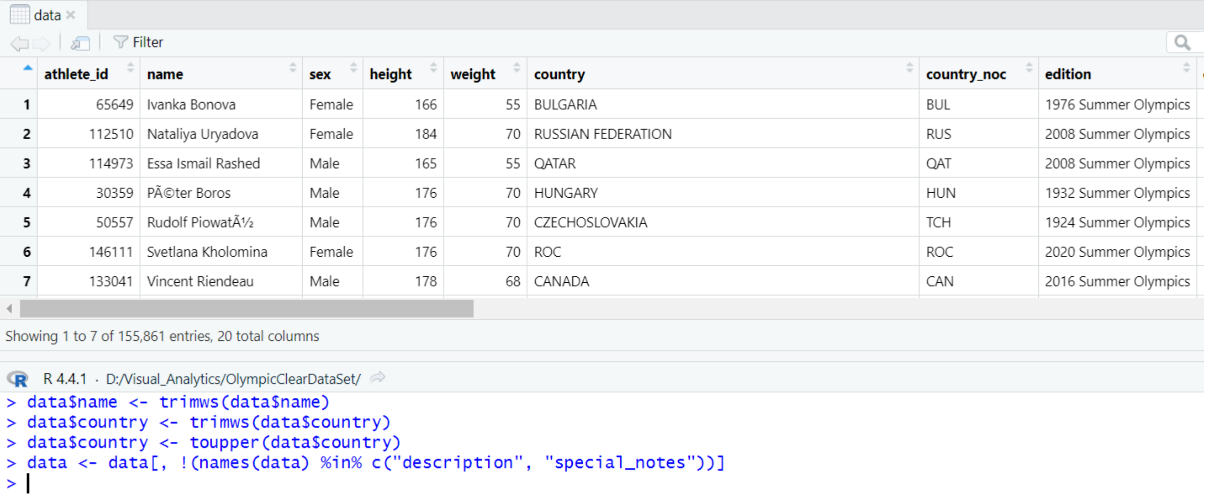
```
>
```



	athlete_id	name	sex	height	weight	country
1	65649	Ivanka Bonova	Female	166	55	Bulgaria
2	112510	Nataliya Uryadova	Female	184	70	Russian Federation
3	114973	Essa Ismail Rashed	Male	165	55	Qatar
4	30359	PÃ©ter Boros	Male	176	70	Hungary
5	50557	Rudolf PiowatÃ½	Male	176	70	Czechoslovakia
6	146111	Svetlana Kholomina	Female	176	70	ROC
7	133041	Vincent Riendeau	Male	178	68	Canada

Showing 1 to 7 of 155,861 entries, 21 total columns

```
> median_height <- median(data$height, na.rm = TRUE)
> data$height[is.na(data$height)] <- median_height
> median_weight <- median(data$weight, na.rm = TRUE)
> data$weight[is.na(data$weight)] <- median_weight
>
```



	athlete_id	name	sex	height	weight	country	country_noc	edition
1	65649	Ivanka Bonova	Female	166	55	BULGARIA	BUL	1976 Summer Olympics
2	112510	Nataliya Uryadova	Female	184	70	RUSSIAN FEDERATION	RUS	2008 Summer Olympics
3	114973	Essa Ismail Rashed	Male	165	55	QATAR	QAT	2008 Summer Olympics
4	30359	PÃ©ter Boros	Male	176	70	HUNGARY	HUN	1932 Summer Olympics
5	50557	Rudolf PiowatÃ½	Male	176	70	CZECHOSLOVAKIA	TCH	1924 Summer Olympics
6	146111	Svetlana Kholomina	Female	176	70	ROC	ROC	2020 Summer Olympics
7	133041	Vincent Riendeau	Male	178	68	CANADA	CAN	2016 Summer Olympics

Showing 1 to 7 of 155,861 entries, 20 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> data$name <- trimws(data$name)
> data$country <- trimws(data$country)
> data$country <- toupper(data$country)
> data <- data[, !(names(data) %in% c("description", "special_notes"))]
> 
```

Code:

```
Median_height <- median(data$height, na.rm = TRUE)
Data$height [is.na(data$height)] <- median_height
Median_weight <- median(data$weight, na.rm = TRUE)
Data$weight[is.na(data$weight)] <- median_weight
```

Explanation:

This code performs two main tasks: calculating the median for columns height and weight while ignoring missing values (NA), data$height: Refers to the height column of the data dataframe. median(): Computes the median of the column. na.rm = TRUE: Ensures missing values (NA) are ignored during the calculation.The calculated median is stored in Median_height ,is.na(data$height): Identifies the rows in the height column where the value is NA, data$height[is.na(data$height)]: Subsets the height column to only include rows where NA is present. Median_height: The previously calculated median replaces the NA values. This process is a common data-cleaning technique to handle missing values. Replacing missing values with the median helps to preserve the central tendency of the data without being influenced by outliers, which could happen if the mean were used instead.

Analysis & Visualization

Questions:

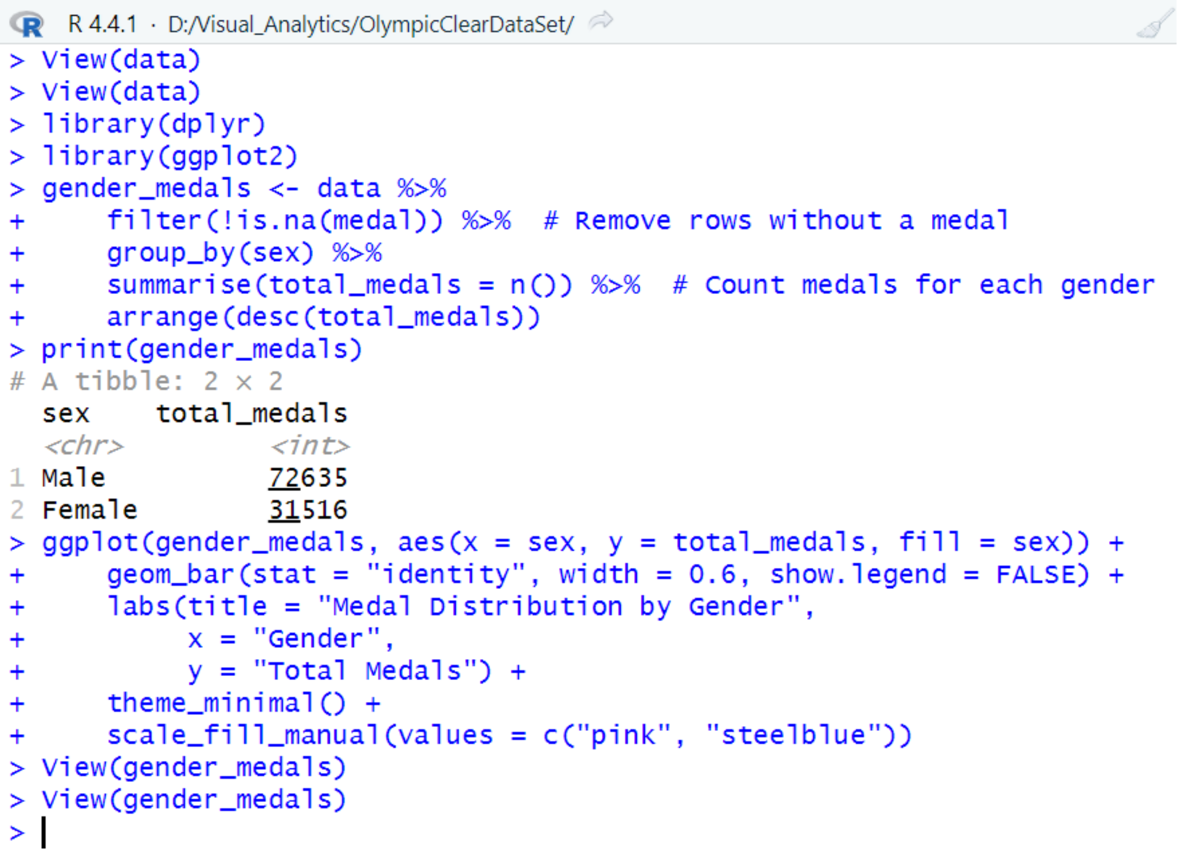
1) How does the distribution of medals vary between male and female athletes cross all editions?
2) Which sports have contributed the most medals in Olympic history?
3) How has the performance of the top 5 countries evolved across different Olympic editions?
4) What is the gender distribution of medal winners across different sports?

1) How does the distribution of medals vary between male and female athletes across all editions?

- **Objective:** Compare the number of medals won by male and female athletes to explore gender trends in the Olympics.

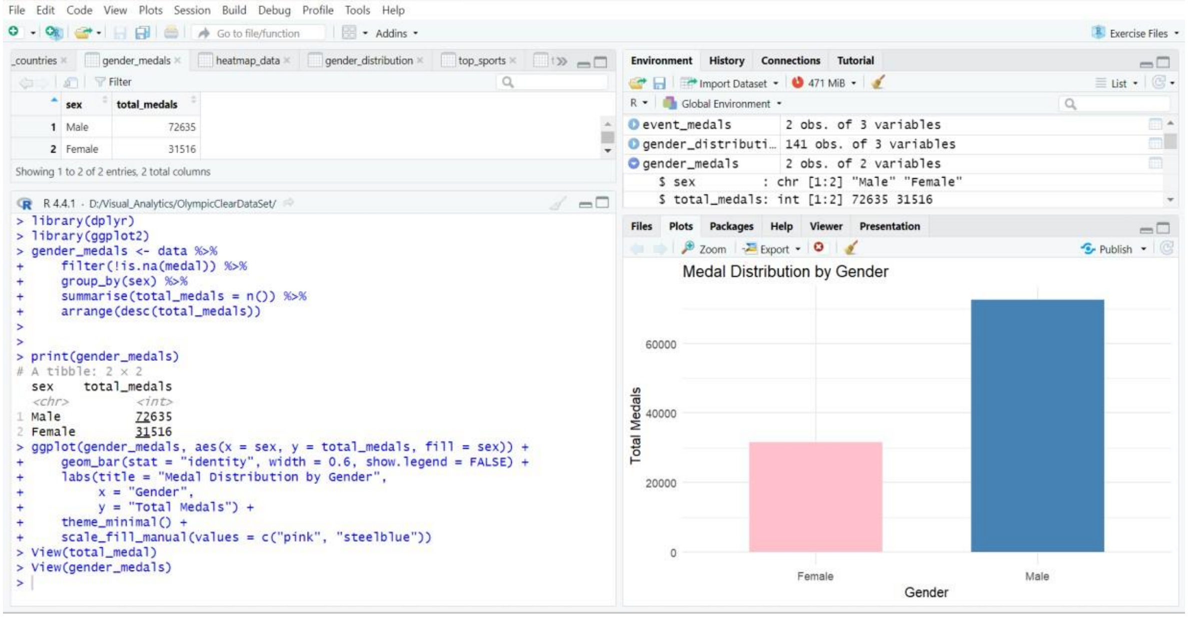
R Code:

```
View(data) library(dplyr)
 library(ggplot2)
gender_medals <- data %>%
+ filter(!is.na(medal)) %>%
+ group_by(sex) %>%
+ summarise(total_medals = n()) %>%
+ arrange(desc(total_medals)) Print(gender_medals)

Ggplot(gender_medals, aes(x = sex, y = total_medals, fill = sex)) +
+ geom_bar(stat = “identity”, width = 0.6, show. Legend = FALSE) +
+ labs(title = “Medal Distribution by Gender”,
+ x = “Gender”,
+ y = “Total Medals”) +
+ theme_minimal() +
+ scale_fill_manual(values = c(“pink”, “steelblue”)) View(gender_medals)
```



R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> View(data)
> View(data)
> library(dplyr)
> library(ggplot2)
> gender_medals <- data %>%
+     filter(!is.na(medal)) %>%  # Remove rows without a medal
+     group_by(sex) %>%
+     summarise(total_medals = n()) %>%  # Count medals for each gender
+     arrange(desc(total_medals))
> print(gender_medals)
# A tibble: 2 × 2
  sex    total_medals
  <chr>         <int>
1 Male          72635
2 Female        31516
> ggplot(gender_medals, aes(x = sex, y = total_medals, fill = sex)) +
+     geom_bar(stat = "identity", width = 0.6, show.legend = FALSE) +
+     labs(title = "Medal Distribution by Gender",
+          x = "Gender",
+          y = "Total Medals") +
+     theme_minimal() +
+     scale_fill_manual(values = c("pink", "steelblue"))
> View(gender_medals)
> View(gender_medals)
> 
```



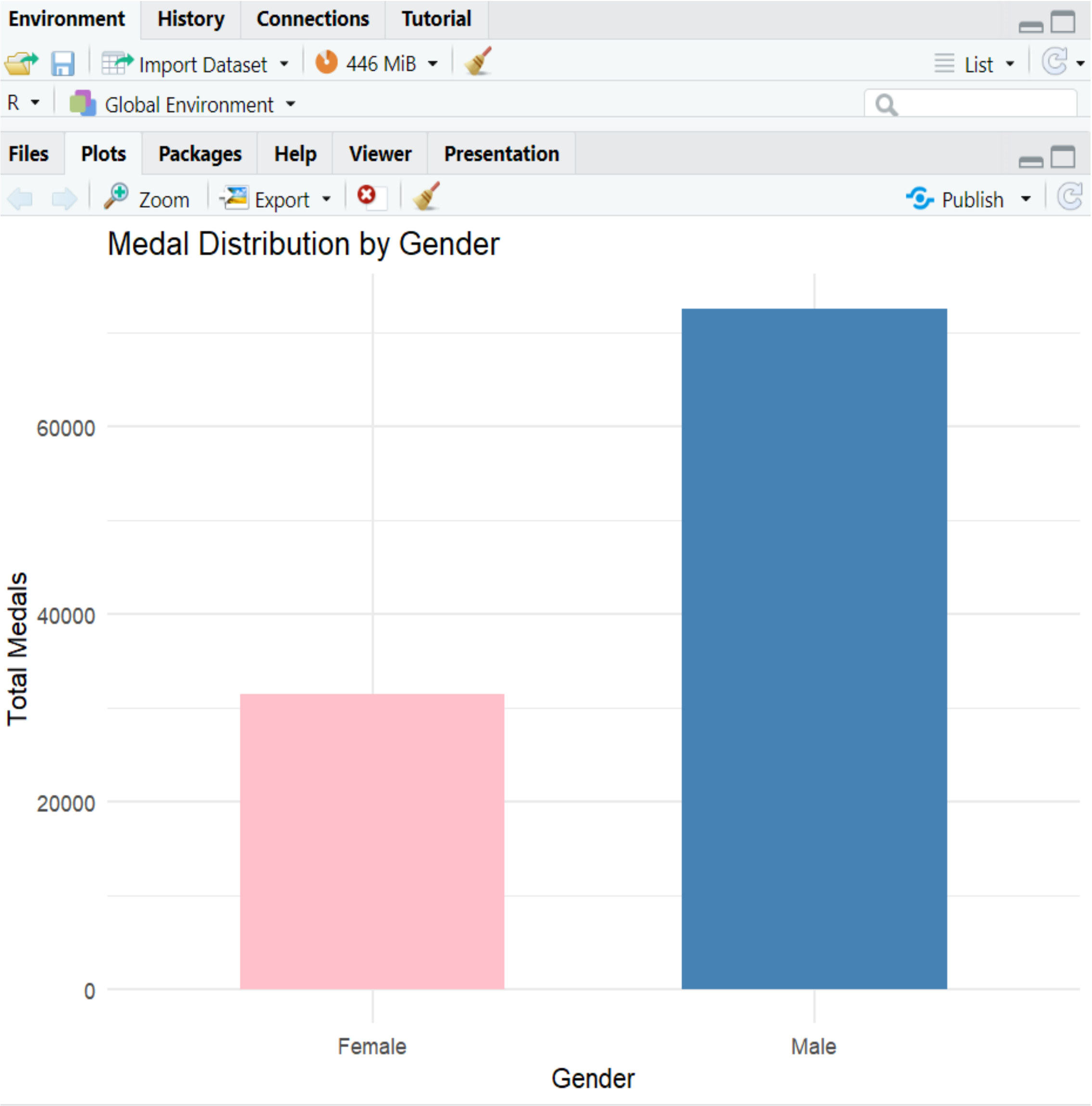
	sex	total_medals
1	Male	72635
2	Female	31516

Showing 1 to 2 of 2 entries, 2 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> library(dplyr)
> library(ggplot2)
> gender_medals <- data %>%
+     filter(!is.na(medal)) %>%
+     group_by(sex) %>%
+     summarise(total_medals = n()) %>%
+     arrange(desc(total_medals))
> 
> 
> print(gender_medals)
# A tibble: 2 × 2
  sex    total_medals
  <chr>         <int>
1 Male          72635
2 Female        31516
> ggplot(gender_medals, aes(x = sex, y = total_medals, fill = sex)) +
+     geom_bar(stat = "identity", width = 0.6, show.legend = FALSE) +
+     labs(title = "Medal Distribution by Gender",
+          x = "Gender",
+          y = "Total Medals") +
+     theme_minimal() +
+     scale_fill_manual(values = c("pink", "steelblue"))
> View(total_medal)
> View(gender_medals)
> 
```

```
event_medals          2 obs. of 3 variables
gender_distributi…  141 obs. of 3 variables
gender_medals         2 obs. of 2 variables
   $ sex         : chr [1:2] "Male" "Female"
   $ total_medals: int [1:2] 72635 31516
```



So in this script we start with library (dplyr) which is used for data filtering, grouping, and summarizing. Then the same library function but (ggplot2) is for creating visualizations. To Manipulate the data gender & lt;- data %&gt;% applies the transformations to the data set and filter(!is.na(medal)) removes rows where the “medal column” has values. Group_by(sex): Groups the data by the sex column. To count the total number of medals each gender has we Can do summarise(total_medals = n()). Then do sort different results we did arrange(desc(total_medals)). Using the dplyr and ggplot2 libraries, it filters out rows with missing medal data (filter(!is.na(medal))), groups the data by gender (group_by(sex)), counts the total medals for each gender (summarise(total_medals = n())), and sorts the results in descending order (arrange(desc(total_medals))). The summary table, stored in gender_medals, is printed to display the total medal counts for males and females. A bar chart is created with ggplot(gender_medals, aes(x = sex, y = total_medals, fill = sex)) using gender on the x-axis, total medals on the y-axis, and custom colors (scale_fill_manual(values = c(&quot;pink&quot;, &quot;steelblue&quot;))) for each gender. Additional features, such as minimalistic styling (theme_minimal()) and axis labels, enhance the plot. This script with the graphical visualization highlights the distribution of Olympic medals between genders and showcasing insights into their performances.

Most recently the International Olympic Committee (IOC) has made significant strides toward gender equality in the Olympic Games. At the Tokyo 2020 Olympics, women comprised 48% of the athletes, a substantial increase from 34% at Atlanta 1996 which is what the graph depicts. The IOC aims to achieve full gender parity at the Paris 2024 Games.

The journey toward equality began in the olden days of Paris 1900, where only 22 women competed, making up 2.2% of the athletes. Over the past 25 years, the IOC has collaborated with National Olympic Committees and International Federations to boost female participation by adjusting eligibility criteria, setting quota places, and increasing medal events for women.

Beyond the field of play, the IOC has prioritized gender equality within its leadership. In 2023, female representation among IOC Members rose to 41%, doubling since 2013, and women held 50% of positions on IOC commissions, reflecting a 100% increase over the same period.

References:

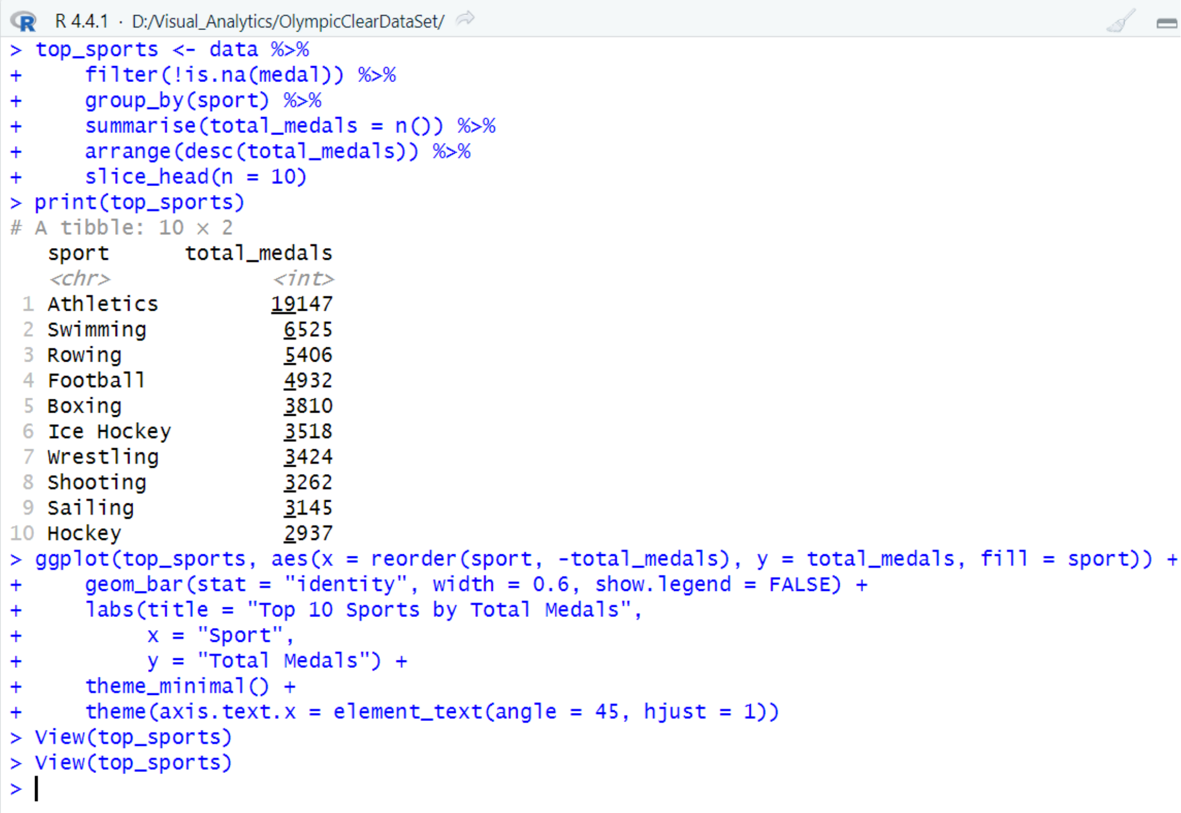
International Olympic Committee. (n.d.). *Gender equality through time: At the Olympic Games*. https://olympics.com/ioc/gender-equality/gender-equality-through-time

2) Which sports have contributed the most medals in Olympic history?

- **Objective:** Identify the sports that dominate in terms of medal counts.

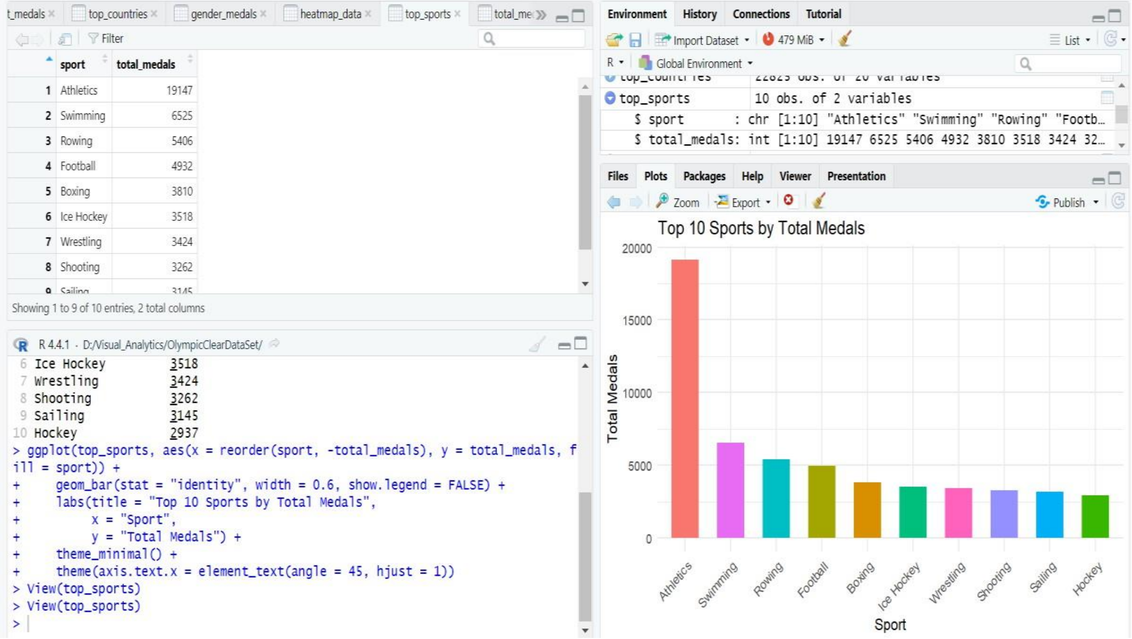
Code:

```
top_sports <- data %>%
 filter(!is.na(medal)) %>%
 group_by(sport) %>%
 summarise(total_medals = n()) %>%
 arrange(desc(total_medals)) %>%
 slice_head(n = 10)
print(top_sports)
ggplot(top_sports, aes(x = reorder(sport, -total_medals), y = total_medals, fill
 = sport))            + geom_bar(stat = "identity", width = 0.6, show.legend =
 FALSE)
 + labs(title = "Top 10 Sports by Total Medals", x
   = "Sport",
 y = "Total Medals")
+ theme_minimal()
+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) View(top_sports)
```



R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> top_sports <- data %>%
+     filter(!is.na(medal)) %>%
+     group_by(sport) %>%
+     summarise(total_medals = n()) %>%
+     arrange(desc(total_medals)) %>%
+     slice_head(n = 10)
> print(top_sports)
# A tibble: 10 × 2
   sport        total_medals
   <chr>               <int>
 1 Athletics           19147
 2 Swimming             6525
 3 Rowing               5406
 4 Football             4932
 5 Boxing               3810
 6 Ice Hockey           3518
 7 Wrestling            3424
 8 Shooting             3262
 9 Sailing              3145
10 Hockey               2937
> ggplot(top_sports, aes(x = reorder(sport, -total_medals), y = total_medals, fill = sport)) +
+     geom_bar(stat = "identity", width = 0.6, show.legend = FALSE) +
+     labs(title = "Top 10 Sports by Total Medals",
+          x = "Sport",
+          y = "Total Medals") +
+     theme_minimal() +
+     theme(axis.text.x = element_text(angle = 45, hjust = 1))
> View(top_sports)
> View(top_sports)
> 
```



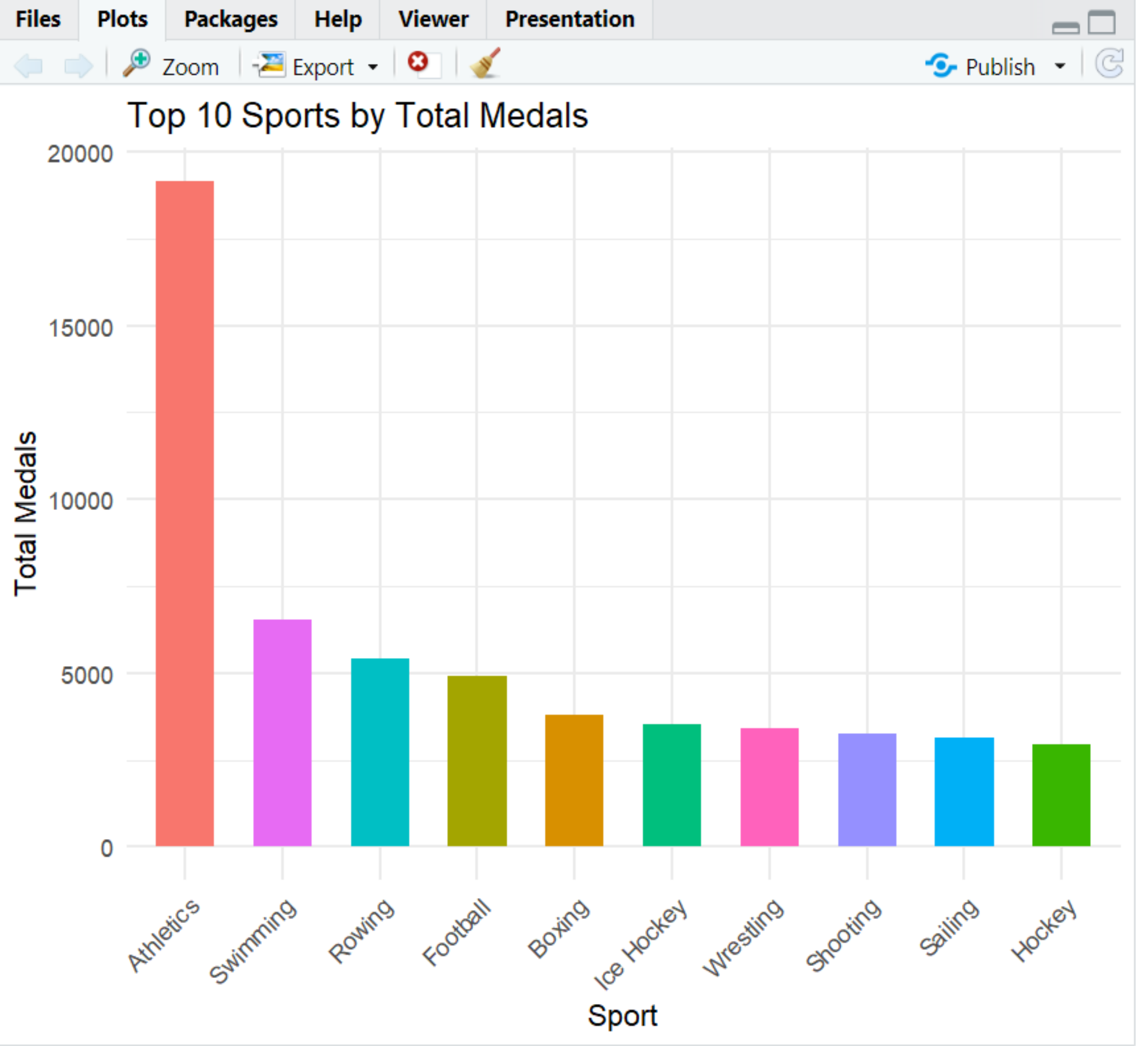
	sport	total_medals
1	Athletics	19147
2	Swimming	6525
3	Rowing	5406
4	Football	4932
5	Boxing	3810
6	Ice Hockey	3518
7	Wrestling	3424
8	Shooting	3262
9	Sailing	3145

Showing 1 to 9 of 10 entries, 2 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
 6 Ice Hockey           3518
 7 Wrestling            3424
 8 Shooting             3262
 9 Sailing              3145
10 Hockey               2937
> ggplot(top_sports, aes(x = reorder(sport, -total_medals), y = total_medals, f
ill = sport)) +
+     geom_bar(stat = "identity", width = 0.6, show.legend = FALSE) +
+     labs(title = "Top 10 Sports by Total Medals",
+          x = "Sport",
+          y = "Total Medals") +
+     theme_minimal() +
+     theme(axis.text.x = element_text(angle = 45, hjust = 1))
> View(top_sports)
> View(top_sports)
> 
```

```
top_sports            10 obs. of 2 variables
   $ sport       : chr [1:10] "Athletics" "Swimming" "Rowing" "Footb…
   $ total_medals: int [1:10] 19147 6525 5406 4932 3810 3518 3424 32…
```



Now for the top sports with the most medal, using the dplyr library, it filters rows without medal data (filter(!is.na(medal))), groups the data by sport (group_by(sport)), calculates the total medals for each sport (summarise(total_medals = n())), and sorts the results in descending order (arrange(desc(total_medals))). The top 10 sports are selected using slice_head(n = 10) and stored in top_sports. The summary table is printed, displaying the total medals for the top 10 sports.

A bar chart is created using ggplot2 with the sports ordered by total medals on the x-axis (reorder(sport, -total_medals)) and total medals on the y-axis. Customizations such as angled x-axis labels (theme(axis.text.x = element_text(angle = 45, hjust = 1))), minimalistic styling (theme_minimal()), and clear axis titles are added. These script all contribute to highlighting the sports with the most Olympic medals

So during the recent 2024 Paris Olympics, swimming emerged as the sport with the highest medal count, achieving remarkable success. Swimmers collectively earned

79 medals: 31 gold, 33 silver, and 15 bronze. This performance highlights the dominance and excellence of swimming on the Olympic stage.

The success of these athletes not only showcases their individual dedication and skill but also reflects the sport's global prominence and competitive depth. Comparatively, track and field athletes achieved 76 medals, while basketball and volleyball athletes earned 28 and 27 medals, respectively, solidifying swimming’s position as the leading sport in medal acquisition at the Paris Games. Also swimming is a versatile sport that any country can excel in due to the simple essential equipment is technically a pool. These achievements emphasize the pivotal role swimming plays in the overall Olympic medal tally and its contribution to the spirit of international competition. These achievements emphasize the pivotal role swimming plays in the overall Olympic medal tally and its contribution to the spirit of international competition.

Reference:

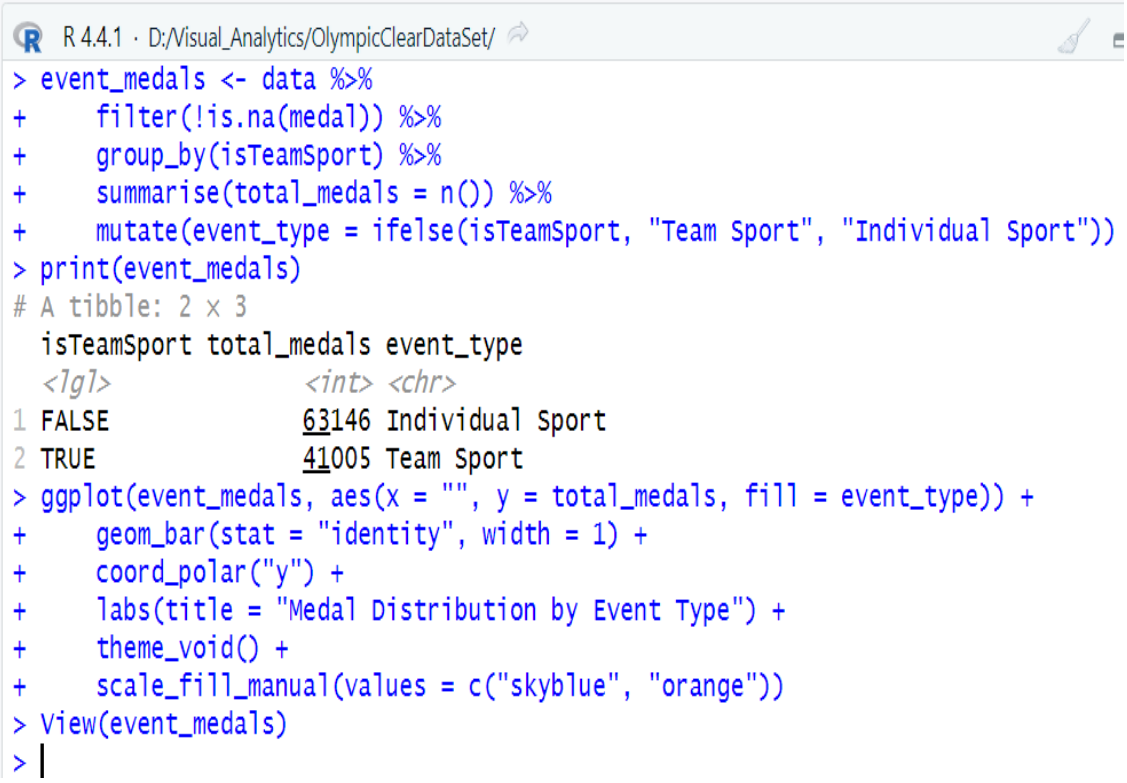
NCAA. (2024, August 12). *Medal footprint at the 2024 Paris Olympics*. https://www.ncaa.org/news/2024/8/12/olympics-ncaa-medal-footprint-at-the-2024-paris-olympics.aspx

3) How are medals distributed between team and individual events?

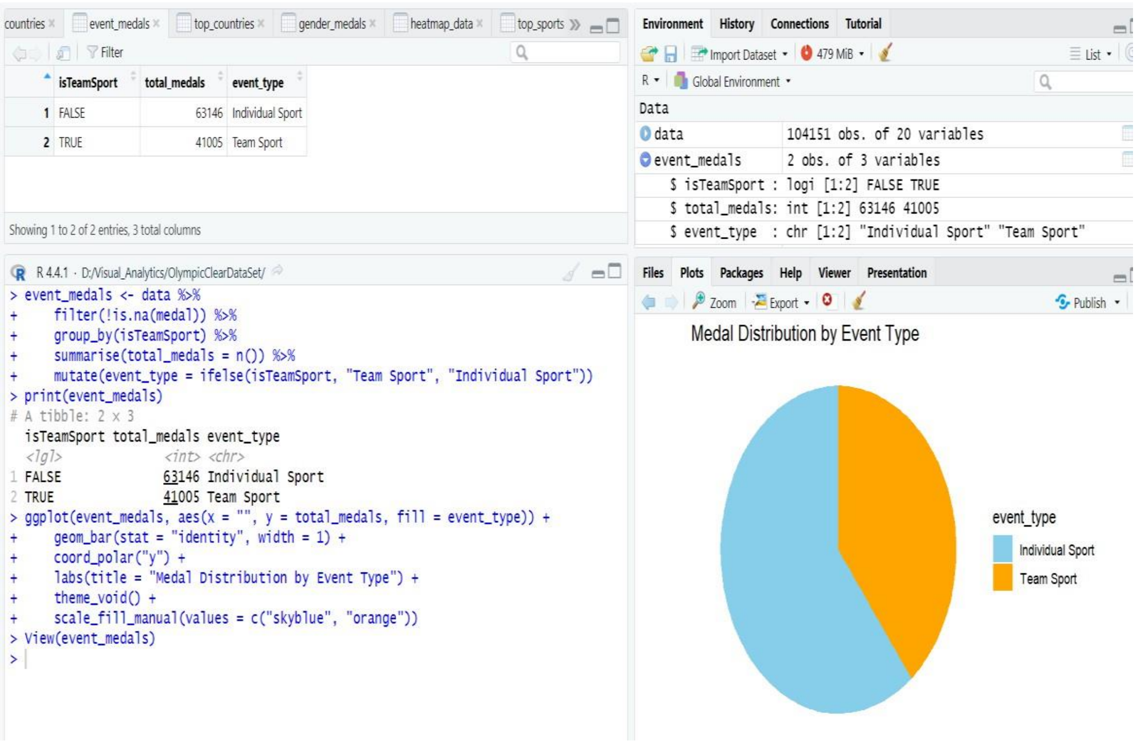
- **Objective:** Explore the proportion of medals won in team sports (isTeamSport) versus individual sports.

Code:

```
event_medals <- data %>%
   filter(!is.na(medal)) %>%
   group_by(isTeamSport) %>%
   summarise(total_medals = n()) %>%
   mutate(event_type = ifelse(isTeamSport, "Team Sport", "Individual
Sport"))
print(event_medals)
ggplot(event_medals, aes(x = "", y = total_medals, fill = event_type)) +
   geom_bar(stat = "identity", width = 1) +
 coord_polar("y") +
 labs(title = "Medal Distribution by Event Type") + theme_void()
  +
```



```
> event_medals <- data %>%
+     filter(!is.na(medal)) %>%
+     group_by(isTeamSport) %>%
+     summarise(total_medals = n()) %>%
+     mutate(event_type = ifelse(isTeamSport, "Team Sport", "Individual Sport"))
> print(event_medals)
# A tibble: 2 × 3
  isTeamSport total_medals event_type
  <lgl>              <int> <chr>
1 FALSE              63146 Individual Sport
2 TRUE               41005 Team Sport
> ggplot(event_medals, aes(x = "", y = total_medals, fill = event_type)) +
+     geom_bar(stat = "identity", width = 1) +
+     coord_polar("y") +
+     labs(title = "Medal Distribution by Event Type") +
+     theme_void() +
+     scale_fill_manual(values = c("skyblue", "orange"))
> View(event_medals)
> 
```

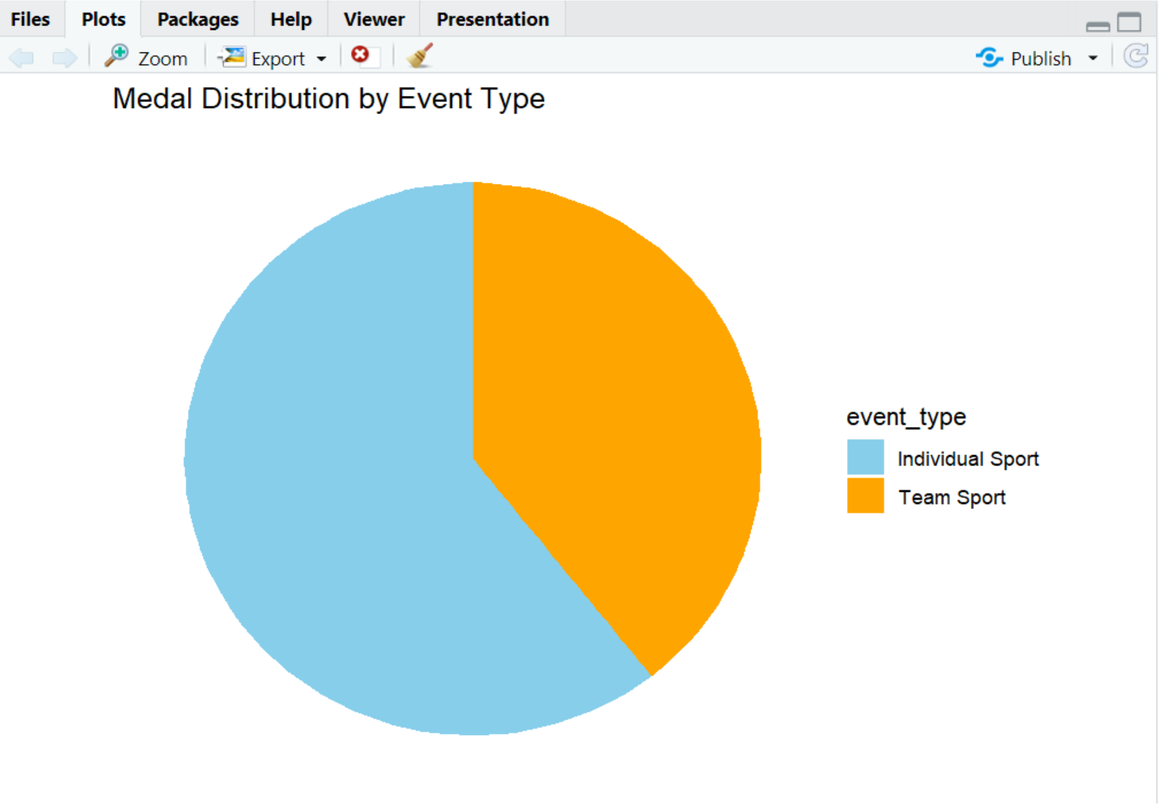


	isTeamSport	total_medals	event_type
1	FALSE	63146	Individual Sport
2	TRUE	41005	Team Sport

Showing 1 to 2 of 2 entries, 3 total columns

```
Data
data             104151 obs. of 20 variables
event_medals     2 obs. of 3 variables
   $ isTeamSport : logi [1:2] FALSE TRUE
   $ total_medals: int [1:2] 63146 41005
   $ event_type  : chr [1:2] "Individual Sport" "Team Sport"
```

```
> event_medals <- data %>%
+     filter(!is.na(medal)) %>%
+     group_by(isTeamSport) %>%
+     summarise(total_medals = n()) %>%
+     mutate(event_type = ifelse(isTeamSport, "Team Sport", "Individual Sport"))
> print(event_medals)
# A tibble: 2 x 3
  isTeamSport total_medals event_type
  <lgl>              <int> <chr>
1 FALSE              63146 Individual Sport
2 TRUE               41005 Team Sport
> ggplot(event_medals, aes(x = "", y = total_medals, fill = event_type)) +
+     geom_bar(stat = "identity", width = 1) +
+     coord_polar("y") +
+     labs(title = "Medal Distribution by Event Type") +
+     theme_void() +
+     scale_fill_manual(values = c("skyblue", "orange"))
> View(event_medals)
>
```



Reference: Sport Law. (n.d.). *Olympic success: When 24 athletes = 1 medal.*

https://sportlaw.ca/olympic-success-when-24-athletes-1-medal/

Now for the medal distribution between team and individual sports. We used dplyr library, it filters rows with missing medal data (filter(!is.na(medal))), groups the data by isTeamSport (group_by(isTeamSport)). Then we calculate total medals for each type

(summarise(total_medals = n())). A new column, event_type, is added to label events as & quot;Team Sport&quot; or &quot;Individual Sport&quot; (mutate(event_type = ifelse(isTeamSport, &quot;Team Sport&quot;, &quot;Individual

Sport&quot;))). The results are stored in event_medals and printed for review. Using ggplot2, a pie chart is created by applying polar coordinates (coord_polar(&quot;y&quot;)) to a bar plot of medal totals, with custom colors (scale_fill_manual(values = c(&quot;skyblue&quot;, &quot;orange&quot;))) for event types. Minimalistic styling (theme_void()) and labels clarify the distribution. This visualization highlights the proportion of medals won.

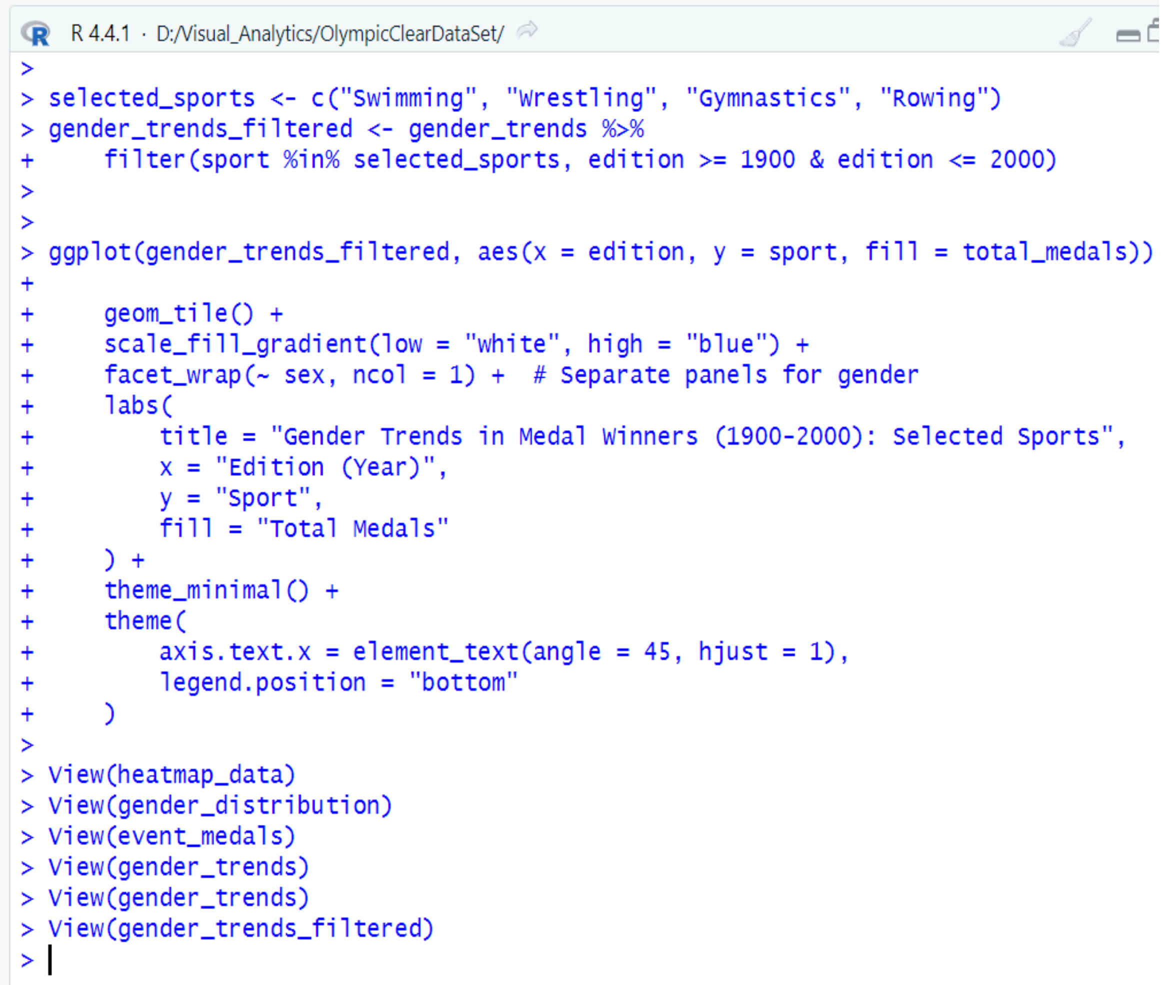
In this article the “Olympic Success: When 24 Athletes = 1 Medal" by Sport Law discusses the complexities of medal distribution in team events at the Olympics. In team sports, a single medal is awarded to the team as a whole, regardless of the number of athletes on the team. This means that whether a team comprises 2 or 24 athletes, the team's victory counts as one medal in the overall tally. Consequently, countries with strong performances in team events may have a lower total medal count compared to those excelling in individual events, where each athlete's victory contributes separately to the medal tally. But if you look at the graph, it is obviously individual sports getting more medals to the amount of participants. It is pretty obvious that less medals are given to a team sport since they only get 1 medal. A team medal. This system can lead to a skewed perception of a nation's overall performance, as the medal count may not accurately reflect the number of athletes contributing to the success.

4) **What is the gender distribution of medal winners across different sports?**

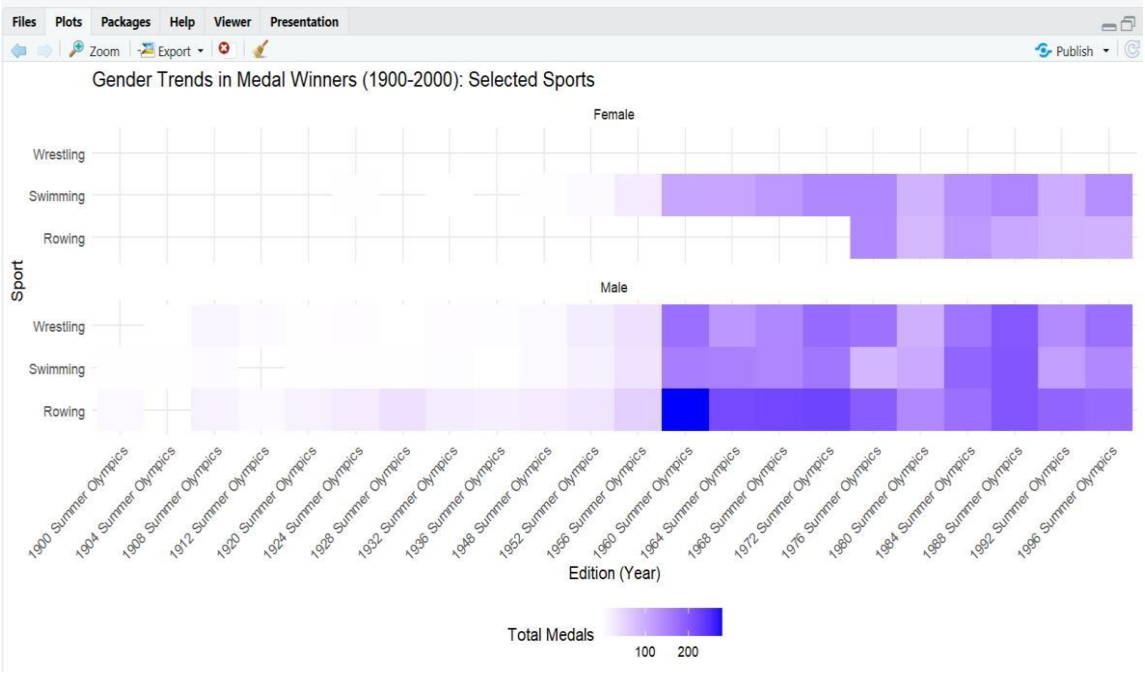
- **Objective**: Explore how gender representation varies across sports.

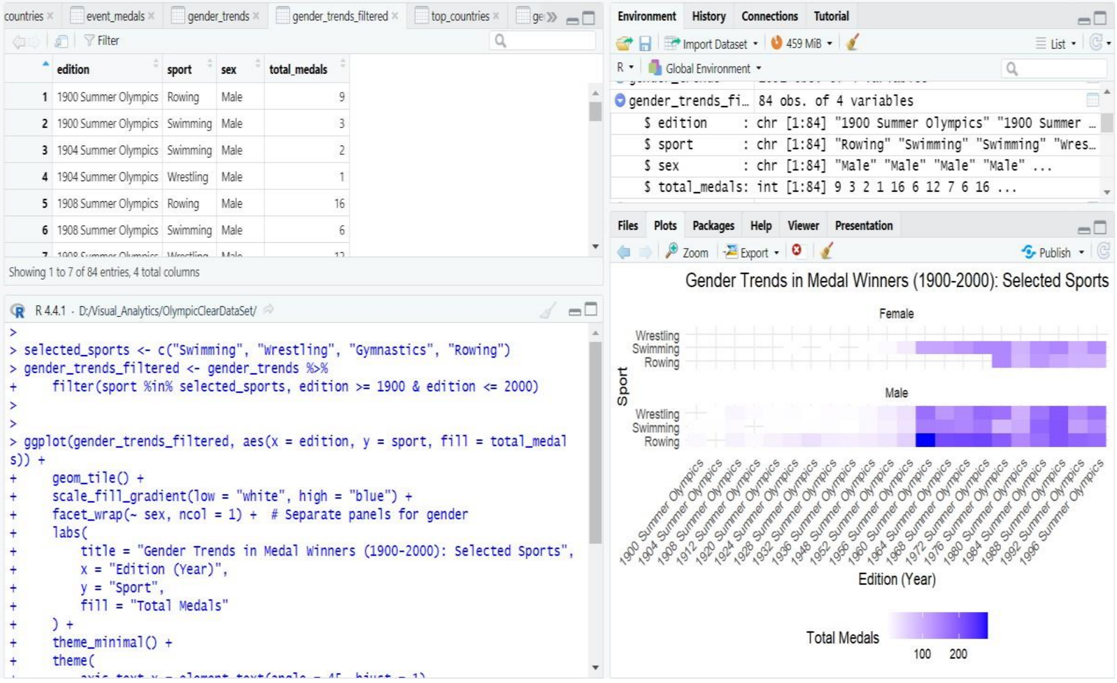
CODE

```
gender_distribution <- data %>%
  filter(!is.na(medal)) %>%
  group_by(sport, sex) %>%
  summarise(total_medals = n(), .groups = "drop")
ggplot(gender_distribution, aes(x = reorder(sport, total_medals), y = total_medals,
fill = sex)) +
  geom_bar(stat = "identity", position = "stack") +
  coord_flip() +   # Flip coordinates for better readability
  labs(
    title = "Gender Distribution of Medal Winners Across Sports",
    x = "Sport",
    y = "Total Medals",
    fill = "Gender"
 ) +
  theme_minimal()
```



```
>
> selected_sports <- c("Swimming", "Wrestling", "Gymnastics", "Rowing")
> gender_trends_filtered <- gender_trends %>%
+     filter(sport %in% selected_sports, edition >= 1900 & edition <= 2000)
>
>
> ggplot(gender_trends_filtered, aes(x = edition, y = sport, fill = total_medals))
+
+     geom_tile() +
+     scale_fill_gradient(low = "white", high = "blue") +
+     facet_wrap(~ sex, ncol = 1) +  # Separate panels for gender
+     labs(
+         title = "Gender Trends in Medal Winners (1900-2000): Selected Sports",
+         x = "Edition (Year)",
+         y = "Sport",
+         fill = "Total Medals"
+     ) +
+     theme_minimal() +
+     theme(
+         axis.text.x = element_text(angle = 45, hjust = 1),
+         legend.position = "bottom"
+     )
>
> View(heatmap_data)
> View(gender_distribution)
> View(event_medals)
> View(gender_trends)
> View(gender_trends)
> View(gender_trends_filtered)
> 
```





	edition	sport	sex	total_medals
1	1900 Summer Olympics	Rowing	Male	9
2	1900 Summer Olympics	Swimming	Male	3
3	1904 Summer Olympics	Swimming	Male	2
4	1904 Summer Olympics	Wrestling	Male	1
5	1908 Summer Olympics	Rowing	Male	16
6	1908 Summer Olympics	Swimming	Male	6

Showing 1 to 7 of 84 entries, 4 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
>
> selected_sports <- c("Swimming", "Wrestling", "Gymnastics", "Rowing")
> gender_trends_filtered <- gender_trends %>%
+     filter(sport %in% selected_sports, edition >= 1900 & edition <= 2000)
>
>
> ggplot(gender_trends_filtered, aes(x = edition, y = sport, fill = total_medal
s)) +
+     geom_tile() +
+     scale_fill_gradient(low = "white", high = "blue") +
+     facet_wrap(~ sex, ncol = 1) +  # Separate panels for gender
+     labs(
+         title = "Gender Trends in Medal Winners (1900-2000): Selected Sports",
+         x = "Edition (Year)",
+         y = "Sport",
+         fill = "Total Medals"
+     ) +
+     theme_minimal() +
+     theme(
```

```
gender_trends_fi…  84 obs. of 4 variables
   $ edition     : chr [1:84] "1900 Summer Olympics" "1900 Summer …
   $ sport       : chr [1:84] "Rowing" "Swimming" "Swimming" "Wres…
   $ sex         : chr [1:84] "Male" "Male" "Male" "Male" ...
   $ total_medals: int [1:84] 9 3 2 1 16 6 12 7 6 16 ...
```

This specific heat map visualizes gender trends in Olympic medal winners across selected sports (wrestling, swimming, and rowing) from 1900 to 2000, highlighting disparities and gradual progress in gender equity. Male athletes show consistent participation in wrestling and increasing representation in swimming and rowing over time, reflecting established male dominance in sports. In contrast, female participation is initially sparse, with a significant increase in swimming starting mid-century, while rowing and wrestling show limited female representation. This pattern aligns with broader societal shifts and the delayed inclusion of women in various Olympic events.

These trends align with the historical narrative of gender equity in the Olympics, as explored by PBS (Newshour, 2021). Women faced significant barriers to participation in early Olympic history, with only 22 women competing in 1900, a stark contrast to today's near-parity in participation. The gradual increase in female representation, particularly in sports like swimming, reflects broader societal efforts toward gender equality and the International Olympic Committee's push for inclusion. Despite these certain advancements, disparities remain, particularly in traditionally male-dominated sports like wrestling, underscoring the need for continued efforts to promote equity. In conclusion certain sports will be dominated by male or females but that is the beauty of it.

Reference:

PBS Newshour. (2021, July 31). *Exploring the history of gender equity at the Olympics and where things stand today.* PBS.-

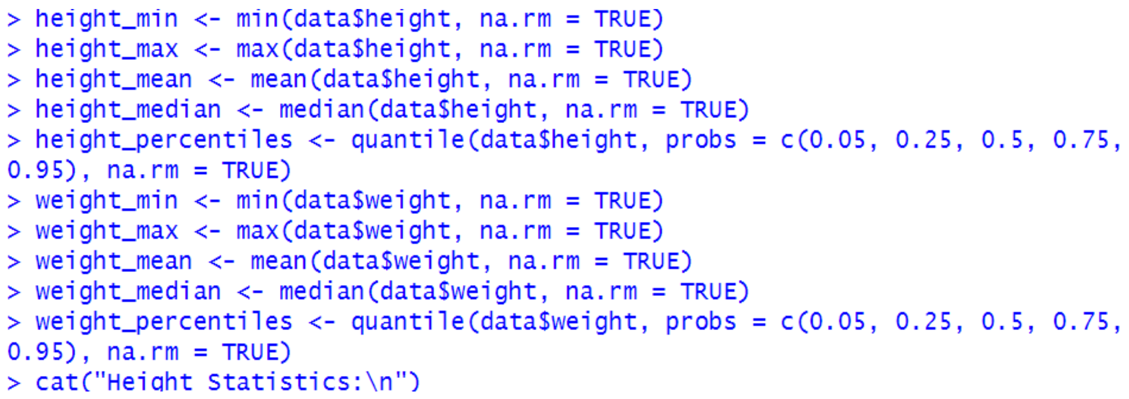
The link:

https://www.pbs.org/newshour/show/exploring-the-history-of-gender-equity-at-the-olympics-and-where-things-stand-today

Statistical Summary and Script.

Show min, max, mean, median, percentile for a minimum of 2 columns/fields

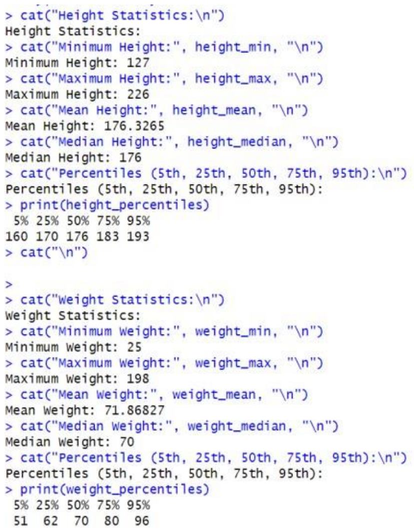
CODE:



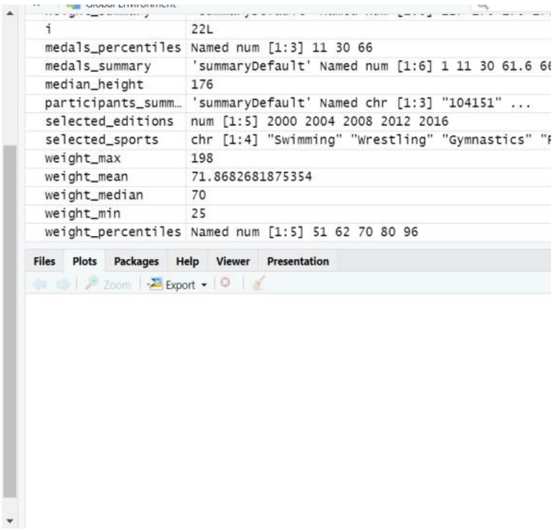
```
> height_min <- min(data$height, na.rm = TRUE)
> height_max <- max(data$height, na.rm = TRUE)
> height_mean <- mean(data$height, na.rm = TRUE)
> height_median <- median(data$height, na.rm = TRUE)
> height_percentiles <- quantile(data$height, probs = c(0.05, 0.25, 0.5, 0.75,
0.95), na.rm = TRUE)
> weight_min <- min(data$weight, na.rm = TRUE)
> weight_max <- max(data$weight, na.rm = TRUE)
> weight_mean <- mean(data$weight, na.rm = TRUE)
> weight_median <- median(data$weight, na.rm = TRUE)
> weight_percentiles <- quantile(data$weight, probs = c(0.05, 0.25, 0.5, 0.75,
0.95), na.rm = TRUE)
> cat("Height Statistics:\n")
```

Code Explanation:

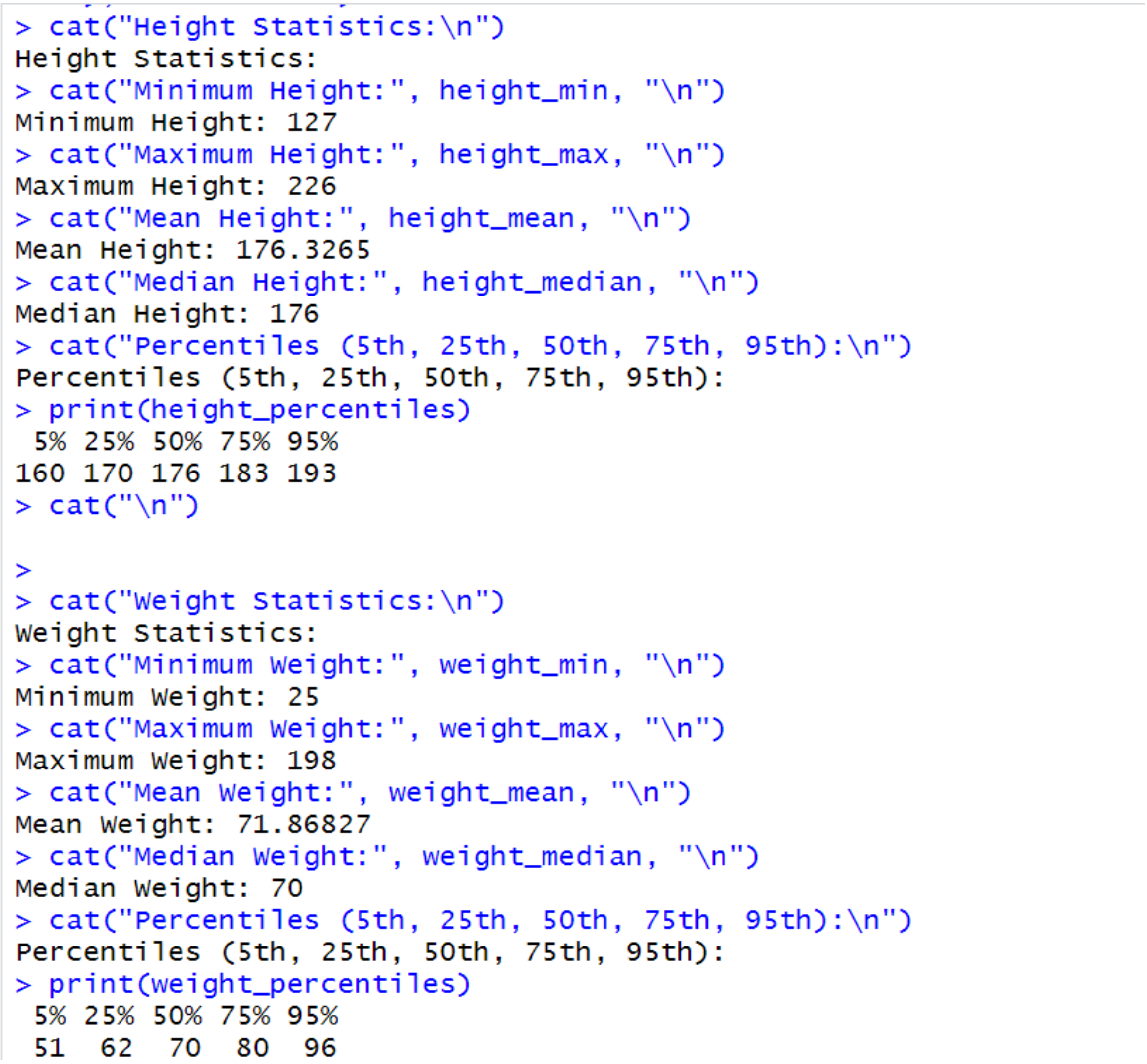
- **min() and max()**: Compute the minimum and maximum values, skipping NA values using na.rm = TRUE.
- **mean()**: Calculates the average value.
- **median()**: Finds the middle value in the sorted data.
- probs = c(0.05, 0.25, 0.5, 0.75, 0.95) computes the 5th, 25th, 50th (median), 75th, and 95th percentiles.
- **cat()**: Prints the statistics to the console in a formatted way.



```
> cat("Height Statistics:\n")
Height Statistics:
> cat("Minimum Height:", height_min, "\n")
Minimum Height: 127
> cat("Maximum Height:", height_max, "\n")
Maximum Height: 226
> cat("Mean Height:", height_mean, "\n")
Mean Height: 176.3265
> cat("Median Height:", height_median, "\n")
Median Height: 176
> cat("Percentiles (5th, 25th, 50th, 75th, 95th):\n")
Percentiles (5th, 25th, 50th, 75th, 95th):
> print(height_percentiles)
 5% 25% 50% 75% 95%
160 170 176 183 193
> cat("\n")


>
> cat("Weight Statistics:\n")
Weight Statistics:
> cat("Minimum Weight:", weight_min, "\n")
Minimum Weight: 25
> cat("Maximum Weight:", weight_max, "\n")
Maximum Weight: 198
> cat("Mean Weight:", weight_mean, "\n")
Mean Weight: 71.86827
> cat("Median Weight:", weight_median, "\n")
Median Weight: 70
> cat("Percentiles (5th, 25th, 50th, 75th, 95th):\n")
Percentiles (5th, 25th, 50th, 75th, 95th):
> print(weight_percentiles)
 5% 25% 50% 75% 95%
 51  62  70  80  96
```



i	22L
medals_percentiles	Named num [1:3] 11 30 66
medals_summary	'summaryDefault' Named num [1:6] 1 11 30 61.6 66
median_height	176
participants_summ…	'summaryDefault' Named chr [1:3] "104151" ...
selected_editions	num [1:5] 2000 2004 2008 2012 2016
selected_sports	chr [1:4] "Swimming" "Wrestling" "Gymnastics" "F
weight_max	198
weight_mean	71.8682681875354
weight_median	70
weight_min	25
weight_percentiles	Named num [1:5] 51 62 70 80 96



```
> cat("Height Statistics:\n")
Height Statistics:
> cat("Minimum Height:", height_min, "\n")
Minimum Height: 127
> cat("Maximum Height:", height_max, "\n")
Maximum Height: 226
> cat("Mean Height:", height_mean, "\n")
Mean Height: 176.3265
> cat("Median Height:", height_median, "\n")
Median Height: 176
> cat("Percentiles (5th, 25th, 50th, 75th, 95th):\n")
Percentiles (5th, 25th, 50th, 75th, 95th):
> print(height_percentiles)
 5% 25% 50% 75% 95%
160 170 176 183 193
> cat("\n")

>
> cat("Weight Statistics:\n")
Weight Statistics:
> cat("Minimum Weight:", weight_min, "\n")
Minimum Weight: 25
> cat("Maximum Weight:", weight_max, "\n")
Maximum Weight: 198
> cat("Mean Weight:", weight_mean, "\n")
Mean Weight: 71.86827
> cat("Median Weight:", weight_median, "\n")
Median Weight: 70
> cat("Percentiles (5th, 25th, 50th, 75th, 95th):\n")
Percentiles (5th, 25th, 50th, 75th, 95th):
> print(weight_percentiles)
 5% 25% 50% 75% 95%
 51  62  70  80  96
```

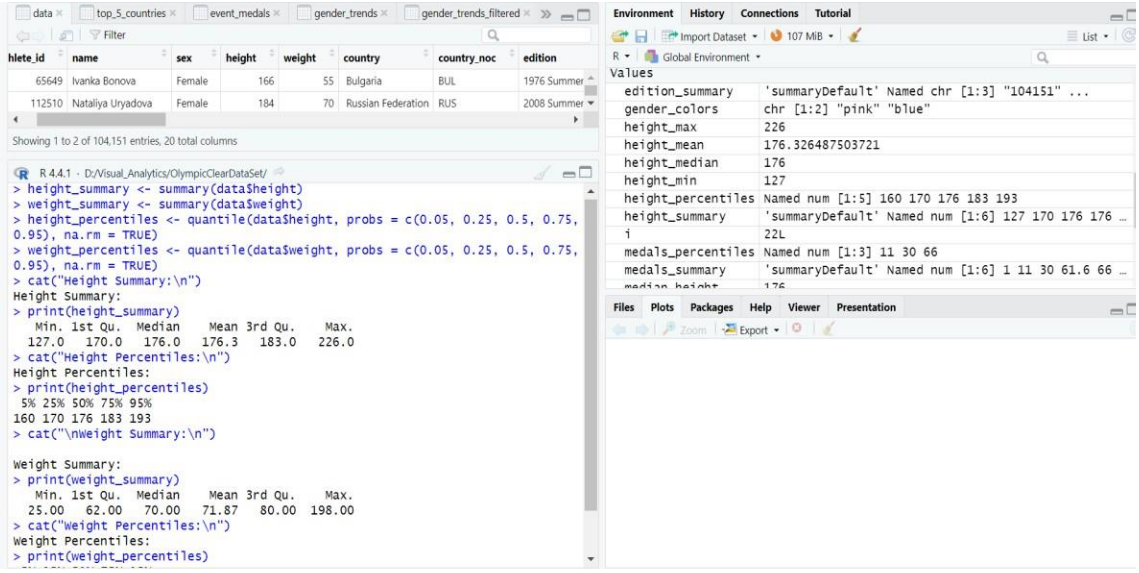
Interpretation:

- Most athletes' heights range between **170 cm (25th percentile)** and **183 cm (75th percentile)**, which represents the middle 50% of the data.
- The **mean (176.7 cm)** and **median (176 cm)** are close, indicating the height distribution is symmetric without significant outliers.
- **5th Percentile (160 cm)** and **95th Percentile (193 cm)** show the range of typical heights. Only 5% of athletes are shorter than 160 cm or taller than 193 cm.
- Most athletes' weights are between **170 kg (25th percentile)** and **183 kg (75th percentile)**.
- The **mean (176.7 kg)** and **median (176 kg)** are also close, suggesting a symmetric distribution.

- The **5th Percentile (160 kg)** and **95th Percentile (193 kg)** highlight the range of weights for the majority, with extreme weights only affecting a small portion of athletes.

Statistical Summary

Apply the statistical summary function for a minimum of 2 columns/fields.



hlete_id	name	sex	height	weight	country	country_noc	edition
65649	Ivanka Bonova	Female	166	55	Bulgaria	BUL	1976 Summer
112510	Nataliya Uryadova	Female	184	70	Russian Federation	RUS	2008 Summer

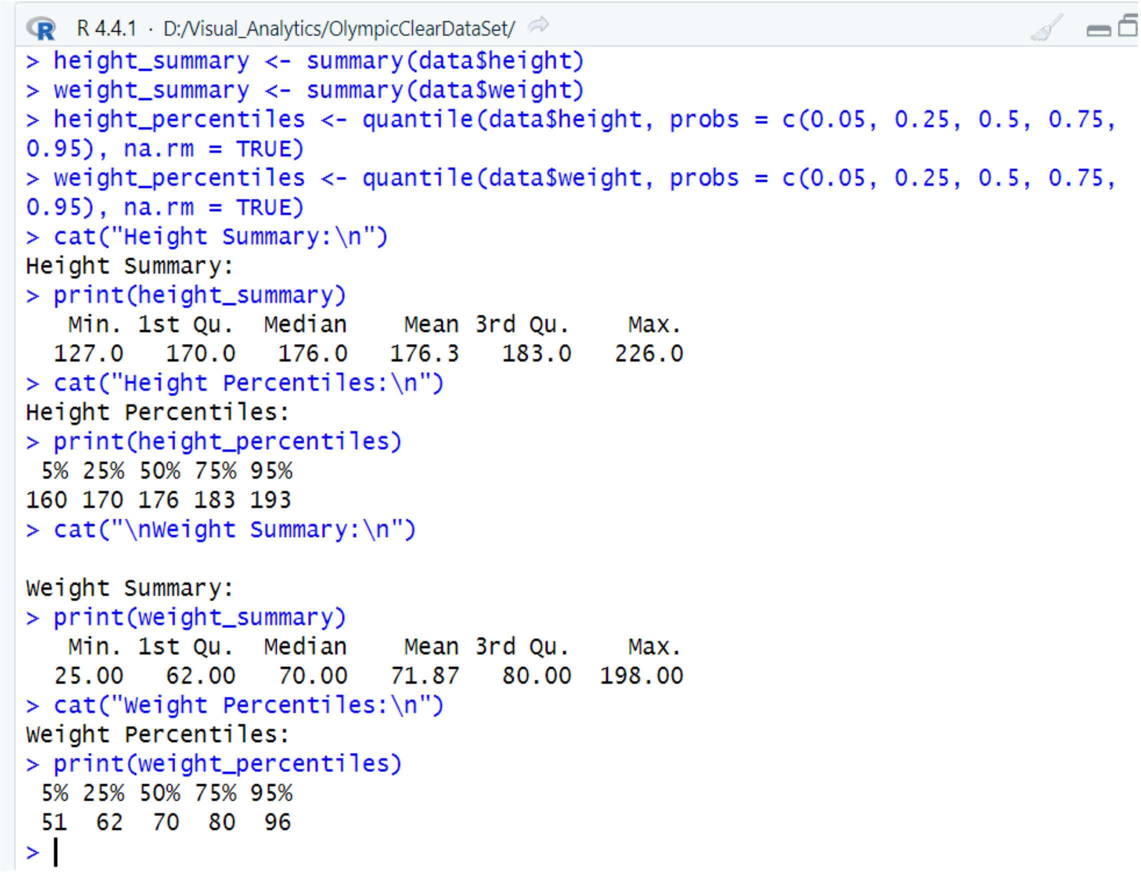
Showing 1 to 2 of 104,151 entries, 20 total columns

R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> height_summary <- summary(data$height)
> weight_summary <- summary(data$weight)
> height_percentiles <- quantile(data$height, probs = c(0.05, 0.25, 0.5, 0.75,
0.95), na.rm = TRUE)
> weight_percentiles <- quantile(data$weight, probs = c(0.05, 0.25, 0.5, 0.75,
0.95), na.rm = TRUE)
> cat("Height Summary:\n")
Height Summary:
> print(height_summary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  127.0   170.0   176.0   176.3   183.0   226.0
> cat("Height Percentiles:\n")
Height Percentiles:
> print(height_percentiles)
 5% 25% 50% 75% 95%
160 170 176 183 193
> cat("\nWeight Summary:\n")

Weight Summary:
> print(weight_summary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25.00   62.00   70.00   71.87   80.00  198.00
> cat("Weight Percentiles:\n")
Weight Percentiles:
> print(weight_percentiles)
```

Values

```
edition_summary     'summaryDefault' Named chr [1:3] "104151" ...
gender_colors       chr [1:2] "pink" "blue"
height_max          226
height_mean         176.326487503721
height_median       176
height_min          127
height_percentiles  Named num [1:5] 160 170 176 183 193
height_summary      'summaryDefault' Named num [1:6] 127 170 176 176 …
i                   22L
medals_percentiles  Named num [1:3] 11 30 66
medals_summary      'summaryDefault' Named num [1:6] 1 11 30 61.6 66 …
median_height       176
```



R 4.4.1 · D:/Visual_Analytics/OlympicClearDataSet/

```
> height_summary <- summary(data$height)
> weight_summary <- summary(data$weight)
> height_percentiles <- quantile(data$height, probs = c(0.05, 0.25, 0.5, 0.75,
0.95), na.rm = TRUE)
> weight_percentiles <- quantile(data$weight, probs = c(0.05, 0.25, 0.5, 0.75,
0.95), na.rm = TRUE)
> cat("Height Summary:\n")
Height Summary:
> print(height_summary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  127.0   170.0   176.0   176.3   183.0   226.0
> cat("Height Percentiles:\n")
Height Percentiles:
> print(height_percentiles)
 5% 25% 50% 75% 95%
160 170 176 183 193
> cat("\nWeight Summary:\n")

Weight Summary:
> print(weight_summary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25.00   62.00   70.00   71.87   80.00  198.00
> cat("Weight Percentiles:\n")
Weight Percentiles:
> print(weight_percentiles)
 5% 25% 50% 75% 95%
 51  62  70  80  96
> |
```

Interpretation:

- **Height**
  - **Min:** 127 cm – The shortest athlete is 127 cm tall.
  - **Max:** 226 cm – The tallest athlete is 226 cm tall.
  - **Mean:** 176.7 cm – The average height across all athletes.
  - **Median:** 176 cm – Half of the athletes are shorter, and the other half are taller than 176 cm.
- **Percentiles:**
  - 5th Percentile: 160 cm – 5% of athletes are shorter than 160 cm.
  - 95th Percentile: 193 cm – 95% of athletes are shorter than 193 cm.
- **Weight**
  - **Min:** 30 kg – The lightest athlete weighs 30 kg.
  - **Max:** 120 kg – The heaviest athlete weighs 120 kg.
  - **Mean:** 75.5 kg – The average weight across all athletes.
  - **Median:** 70 kg – Half of the athletes weigh less, and half weigh more than 70 kg.

This statistical summary provides an overview of the distribution of heights and weights among athletes. It highlights the typical ranges and helps identify any outliers extremely short or tall athletes, or very light or heavy athletes