Data cleaning and data cleansing play a crucial role in the initial stages of data preparation for any data analysis or modeling project. These procedures allow us to verify the accuracy and reliability of the data, leading to more precise and insightful outcomes in the analyses.

PROJECT OBJECTIVES

The objective of this project is to clean the data provided from Kaggle and be made available for analysis.

ABOUT DATASET

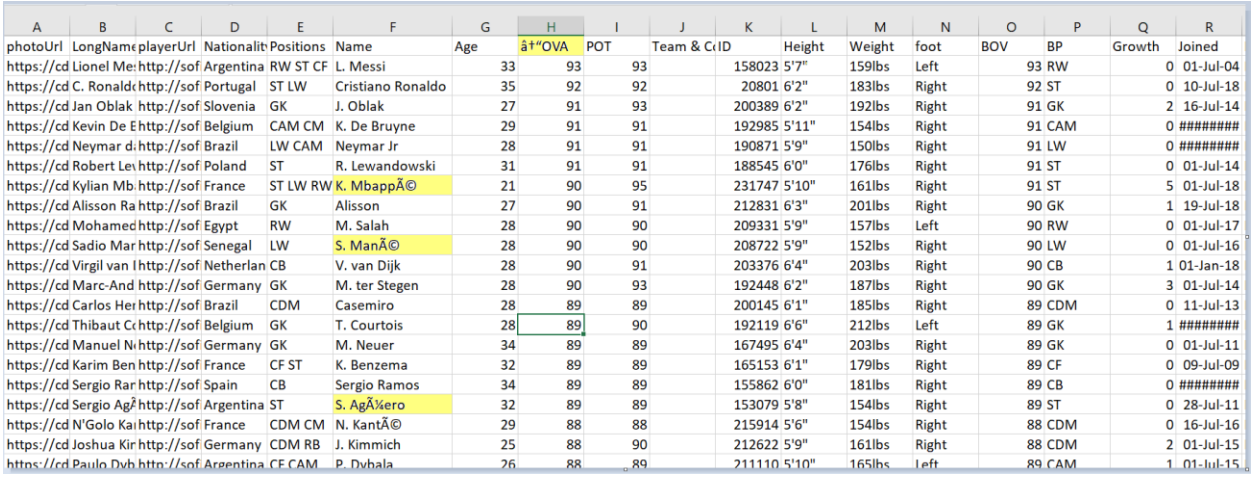
The FIFA 2021 dataset was originally gotten from Kaggle and can be accessed from https://www.kaggle.com/datasets/yagunnersya/fifa-21-messy-raw-dataset-for-cleaning-exploring. The dataset contains information about 18,979 football players and 77 columns of the players statistics and demography in 2021.

Problems To Look Out For In The Data :

Incorrect data types, Null entries, Missing values, Duplicate entries, Errors in spellings and values, Wrong calculations across rows and columns and Irrelevant data.

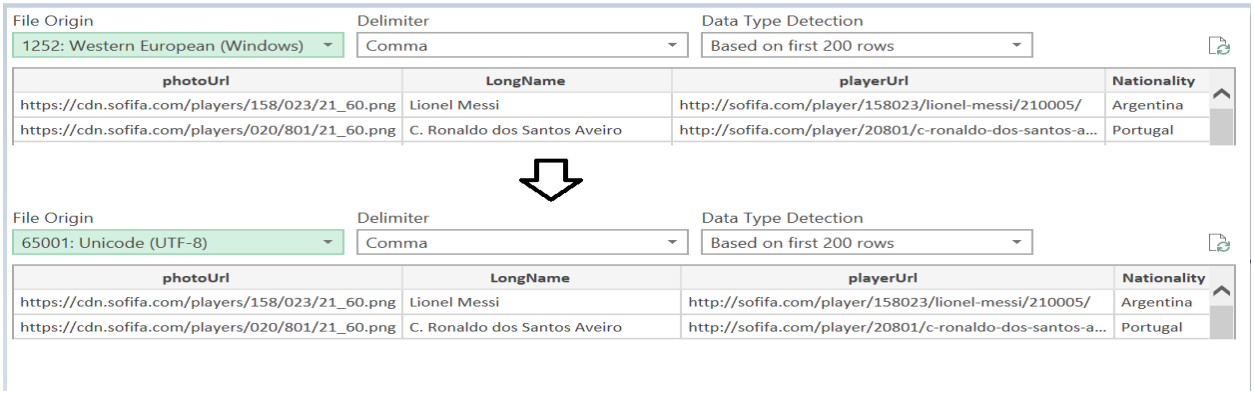
DATA CLEANING AND TRANSFORMATION APPROACH

- The raw data set was downloaded initially as a zipped file which was extracted as csv file the loaded to Microsoft Excel. Afterwards, The data collection has undergone a quick screening to identify any irregularities.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
photoUrl	LongName	playerUrl	Nationalit	Positions	Name	Age	â†"OVA	POT	Team & C	ID	Height	Weight	foot	BOV	BP	Growth	Joined
https://cd	Lionel Me	http://sof	Argentina	RW ST CF	L. Messi	33	93	93		158023	5'7"	159lbs	Left	93	RW	0	01-Jul-04
https://cd	C. Ronald	http://sof	Portugal	ST LW	Cristiano Ronaldo	35	92	92		20801	6'2"	183lbs	Right	92	ST	0	10-Jul-18
https://cd	Jan Oblak	http://sof	Slovenia	GK	J. Oblak	27	91	93		200389	6'2"	192lbs	Right	91	GK	2	16-Jul-14
https://cd	Kevin De B	http://sof	Belgium	CAM CM	K. De Bruyne	29	91	91		192985	5'11"	154lbs	Right	91	CAM	0	########
https://cd	Neymar da	http://sof	Brazil	LW CAM	Neymar Jr	28	91	91		190871	5'9"	150lbs	Right	91	LW	0	########
https://cd	Robert Le	http://sof	Poland	ST	R. Lewandowski	31	91	91		188545	6'0"	176lbs	Right	91	ST	0	01-Jul-14
https://cd	Kylian Mb	http://sof	France	ST LW RW	K. MbappÃ©	21	90	95		231747	5'10"	161lbs	Right	91	ST	5	01-Jul-18
https://cd	Alisson Ra	http://sof	Brazil	GK	Alisson	27	90	91		212831	6'3"	201lbs	Right	90	GK	1	19-Jul-18
https://cd	Mohamed	http://sof	Egypt	RW	M. Salah	28	90	90		209331	5'9"	157lbs	Left	90	RW	0	01-Jul-17
https://cd	Sadio Mar	http://sof	Senegal	LW	S. ManÃ©	28	90	90		208722	5'9"	152lbs	Right	90	LW	0	01-Jul-16
https://cd	Virgil van I	http://sof	Netherlan	CB	V. van Dijk	28	90	91		203376	6'4"	203lbs	Right	90	CB	1	01-Jan-18
https://cd	Marc-And	http://sof	Germany	GK	M. ter Stegen	28	90	93		192448	6'2"	187lbs	Right	90	GK	3	01-Jul-14
https://cd	Carlos Her	http://sof	Brazil	CDM	Casemiro	28	89	89		200145	6'1"	185lbs	Right	89	CDM	0	11-Jul-13
https://cd	Thibaut Cc	http://sof	Belgium	GK	T. Courtois	28	89	90		192119	6'6"	212lbs	Left	89	GK	1	########
https://cd	Manuel Ne	http://sof	Germany	GK	M. Neuer	34	89	89		167495	6'4"	203lbs	Right	89	GK	0	01-Jul-11
https://cd	Karim Ben	http://sof	France	CF ST	K. Benzema	32	89	89		165153	6'1"	179lbs	Right	89	CF	0	09-Jul-09
https://cd	Sergio Rar	http://sof	Spain	CB	Sergio Ramos	34	89	89		155862	6'0"	181lbs	Right	89	CB	0	########
https://cd	Sergio Agâ	http://sof	Argentina	ST	S. AgÃ¼ero	32	89	89		153079	5'8"	154lbs	Right	89	ST	0	28-Jul-11
https://cd	N'Golo Kaı	http://sof	France	CDM CM	N. KantÃ©	29	88	88		215914	5'6"	154lbs	Right	88	CDM	0	16-Jul-16
https://cd	Joshua Kir	http://sof	Germany	CDM RB	J. Kimmich	25	88	90		212622	5'9"	161lbs	Right	88	CDM	2	01-Jul-15
https://cd	Paulo Dyb	http://sof	Argentina	CF CAM	P. Dybala	26	88	89		211110	5'10"	165lbs	Left	89	CAM	1	01-Jul-15

- Subsequently, The dataset loaded into Power Query editor for the data cleaning procedures and the **file origin** has to be changed to UTF-8. It's helping to detect and replace special characters into the proper characters.

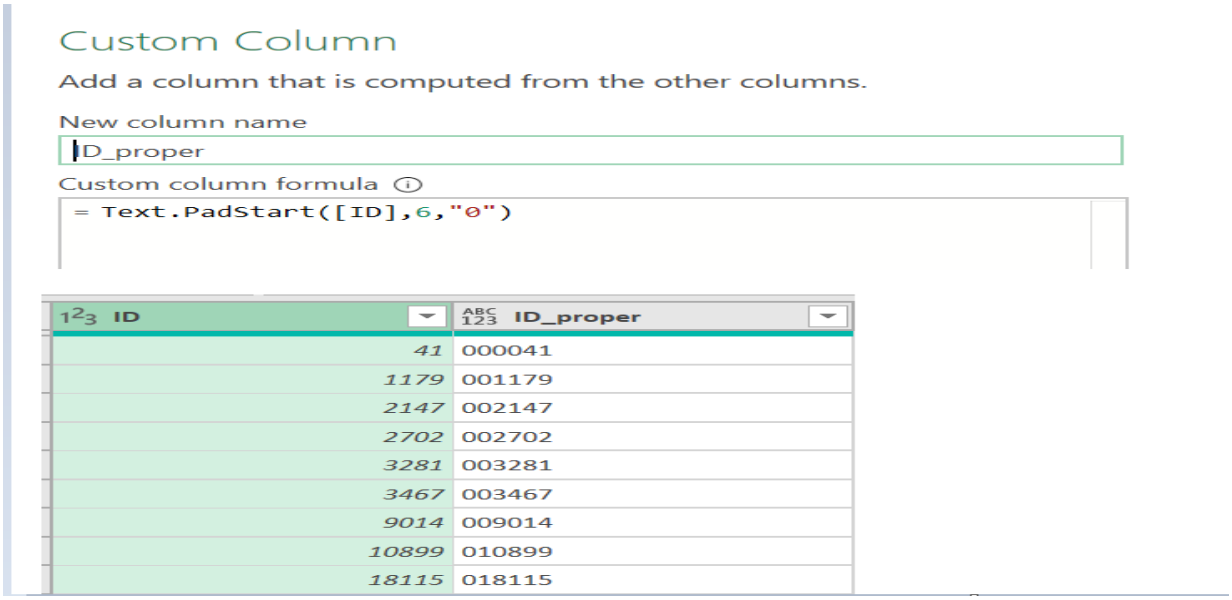


as shown below:

After the importation into Power Query Editor, most of the records in the column has some extra spaces in between them. To handle this scenario, **Trim** function in the **Transform** can be used to remove any trailing spaces in the columns.

Once the extra spaces were eliminated, the data cleaning process commenced, starting from the first column, which contained the ID, and progressed through to the final column. In the following sections, I will outline the procedures employed to rectify any inaccuracies present in each column.

ID: Each record in the table is uniquely identified by this column. Although the records are in numerical format, the data type needs to be changed from number to text to prevent any calculations from being performed using this column. Additionally, since the records in this column have varying text lengths, I have introduced a custom column to add leading zeros, ensuring uniformity in the records.



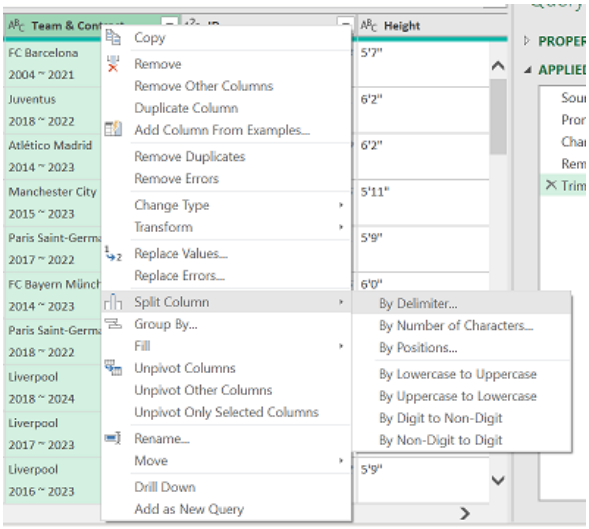
ID	ID_proper
41	000041
1179	001179
2147	002147
2702	002702
3281	003281
3467	003467
9014	009014
10899	010899
18115	018115

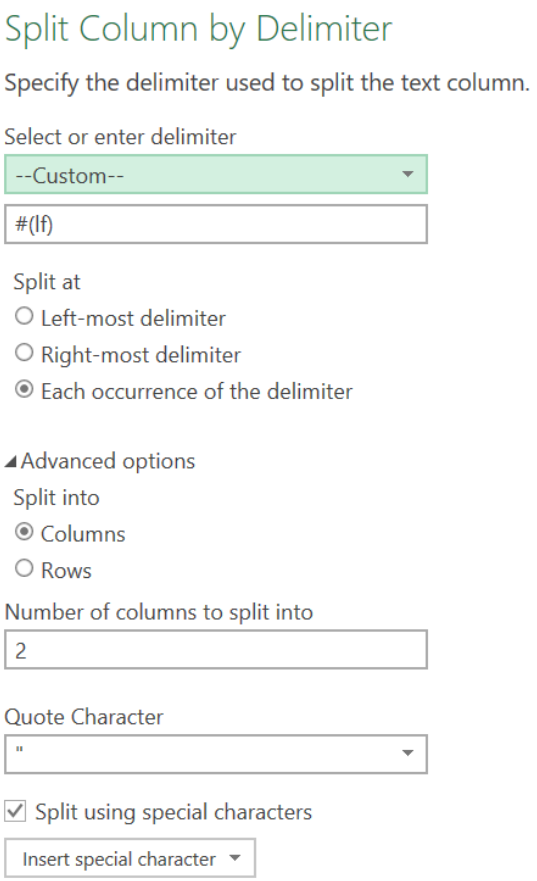
Team & Contract : This particular column holds information about where the player is currently playing, as well as details regarding start and end year of the contract. However, However, it's worth noting that some players have paid contracts, some are on loan, and others are free agents.

Team & Contract
FC Barcelona 2004 ~ 2021
Juventus 2018 ~ 2022
Atlético Madrid 2014 ~ 2023
Manchester City 2015 ~ 2023
Paris Saint-Germain 2017 ~ 2022
FC Bayern München 2014 ~ 2023
Paris Saint-Germain 2018 ~ 2022
Liverpool 2018 ~ 2024

To be precise for further analysis, the column should be transformed into more proper and clear form as shown below and it's approaching steps.

a. Splitting the column by the delimiter and then choose custom





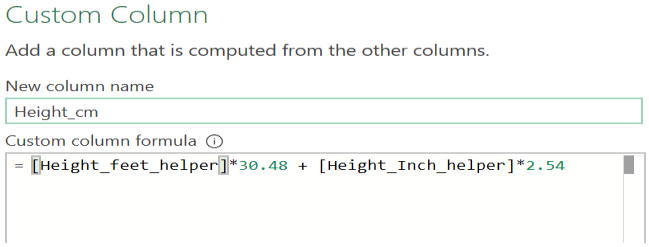
b. The column split into 2 columns and rename each column

Team_name	Contract_in
Paris Saint-Germain	2018 ~ 2022
Liverpool	2018 ~ 2024
Liverpool	2017 ~ 2023
Liverpool	2016 ~ 2023
Liverpool	2018 ~ 2023
FC Barcelona	2014 ~ 2022
Real Madrid	2013 ~ 2023
Real Madrid	2018 ~ 2024
FC Bayern München	2011 ~ 2023
Real Madrid	2009 ~ 2022
Real Madrid	2005 ~ 2021
Manchester City	2011 ~ 2021

c. Moving to next approach, the contract column was broken down into more granularity details by applying combination of **Replace value** and **Split Column** as shown in the following

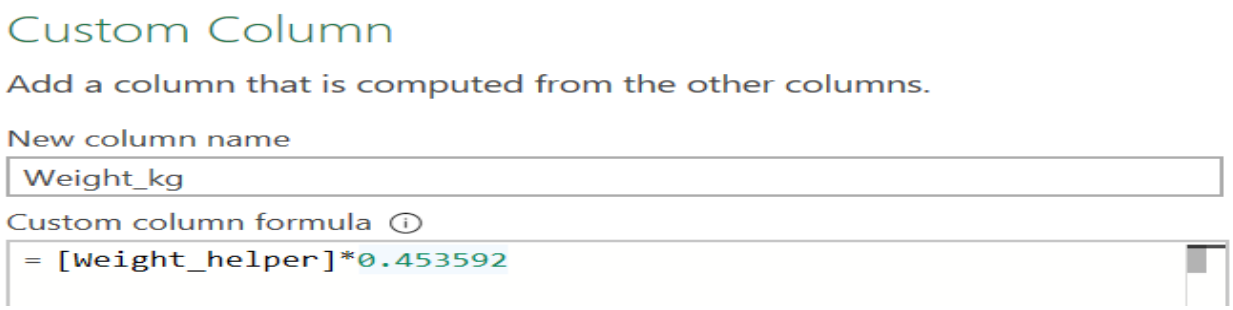
Contract	Start_year_contract	End_year_contract	Contract_duration_year	Agreement
2004 - 2021	2004	2021	17	Contract
2018 - 2022	2018	2022	4	Contract
2014 - 2023	2014	2023	9	Contract
2015 - 2023	2015	2023	8	Contract
2017 - 2022	2017	2022	5	Contract
2014 - 2023	2014	2023	9	Contract
2018 - 2022	2018	2022	4	Contract
2018 - 2024	2018	2024	6	Contract
2017 - 2023	2017	2023	6	Contract

Height : The column contains the height of each player, initially recorded in feet and inches. To enhance precision for further analysis, the measurements were converted into centimeters. The chosen approach for converting the height to centimeters involved splitting the column and extracting the numerical values in feet (multiplied by 30.48) and inches (multiplied by 2.54) to obtain the corresponding measurement in centimeters.



Height_feet	Height_cm
6'2"	187.96
6'2"	187.96
5'11"	180.34
5'9"	175.26
6'0"	182.88
5'10"	177.8
6'3"	190.5
5'9"	175.26
5'9"	175.26
6'4"	193.04
6'2"	187.96
6'1"	185.42
6'6"	198.12
6'4"	193.04
6'1"	185.42

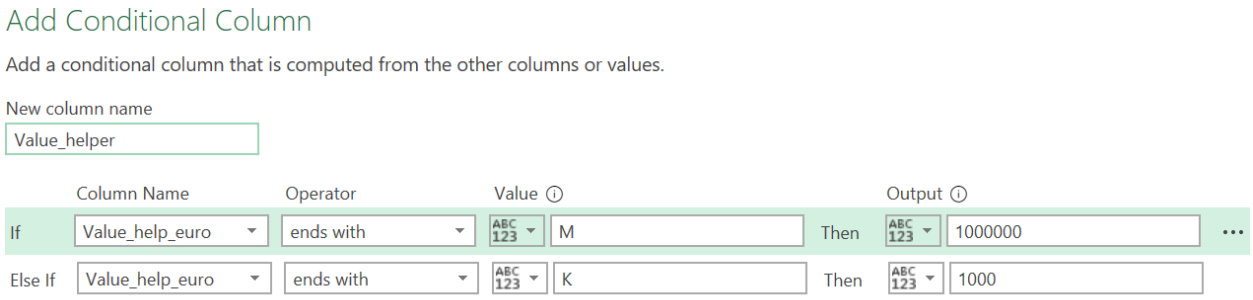
Weight : The column includes the weight of each player, initially recorded in pounds (lbs), which was later converted into kilograms (kg). The chosen method for this conversion involved splitting the column and extracting the values in pounds (multiplied by 0.453592) to obtain the corresponding measurements in kilograms.

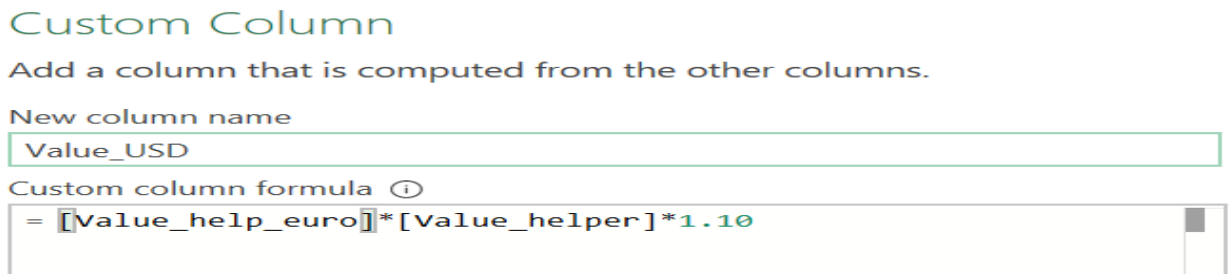


Weight_lbs	Weight_kg
159lbs	72.12
183lbs	83.01
192lbs	87.09
154lbs	69.85
150lbs	68.04
176lbs	79.83
161lbs	73.03
201lbs	91.17
157lbs	71.21
152lbs	68.95
203lbs	92.08

The Value, Wage, and Release-Clause: This column displays the market value, weekly wage, and release clause of players in FIFA 2021. Some entries are indicated with 'K', while others with 'M'. The 'M' values were multiplied by 1,000,000, and the 'K' values by 1,000 using a conditional statement. After removing the "M", "K", and "€" symbols, they were ready for multiplication with the Conditional Column.

Value	Value_help_euro	Value_helper	Value_USD
€67.5M	67.5	1000000	74,250,000.00
€46M	46	1000000	50,600,000.00
€75M	75	1000000	82,500,000.00
€87M	87	1000000	95,700,000.00
€90M	90	1000000	99,000,000.00
€80M	80	1000000	88,000,000.00
€105.5M	105.5	1000000	116,050,000.00
€62.5M	62.5	1000000	68,750,000.00
€78M	78	1000000	85,800,000.00





Each value for **Wage** and **Release_Clause** repeats the precedent steps

Wage	Wage_helper	Wage_USD
€560K	560	616,000.00
€220K	220	242,000.00
€125K	125	137,500.00
€370K	370	407,000.00
€270K	270	297,000.00
€240K	240	264,000.00
€160K	160	176,000.00
€160K	160	176,000.00
€250K	250	275,000.00

Release Clause	Release Clause_helper	Release_helper	Release_clause_USD
€138.4M	138.4	1000000	152,240,000.00
€75.9M	75.9	1000000	83,490,000.00
€159.4M	159.4	1000000	175,340,000.00
€161M	161	1000000	177,100,000.00
€166.5M	166.5	1000000	183,150,000.00
€132M	132	1000000	145,200,000.00
€203.1M	203.1	1000000	223,410,000.00
€120.3M	120.3	1000000	132,330,000.00

Having performed all the essential data cleaning steps, our dataset is now free from errors and prepared for loading into Excel sheets

Dataset view before cleaning

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	photoUrl	LongName	playerUrl	Nationalit	Positions	Name	Age	â†‘OVA	POT	Team & Cc	ID	Height	Weight	foot	BOV	BP	Growth	Joined	Loan Date	Value	Wage	Release Cl	Attacking
2	https://cdi	Lionel Messi	http://sofi	Argentina	RW ST CF	L. Messi	33	93	93		158023	5'7"	159lbs	Left	93	RW	0	01-Jul-04	N/A	â‚¬67.5M	â‚¬560K	â‚¬138.4M	429
3	https://cdi	C. Ronaldo dos S	http://sofi	Portugal	ST LW	Cristiano Ronaldo	35	92	92		20801	6'2"	183lbs	Right	92	ST	0	10-Jul-18	N/A	â‚¬46M	â‚¬220K	â‚¬75.9M	437
4	https://cdi	Jan Oblak	http://sofi	Slovenia	GK	J. Oblak	27	91	93		200389	6'2"	192lbs	Right	91	GK	2	16-Jul-14	N/A	â‚¬75M	â‚¬125K	â‚¬159.4M	95
5	https://cdi	Kevin De Bruyne	http://sofi	Belgium	CAM CM	K. De Bruyne	29	91	91		192985	5'11"	154lbs	Right	91	CAM	0	30-Aug-15	N/A	â‚¬87M	â‚¬370K	â‚¬161M	407
6	https://cdi	Neymar da Silva	http://sofi	Brazil	LW CAM	Neymar Jr	28	91	91		190871	5'9"	150lbs	Right	91	LW	0	03-Aug-17	N/A	â‚¬90M	â‚¬270K	â‚¬166.5M	408
7	https://cdi	Robert Lewando	http://sofi	Poland	ST	R. Lewandowski	31	91	91		188545	6'0"	176lbs	Right	91	ST	0	01-Jul-14	N/A	â‚¬80M	â‚¬240K	â‚¬132M	423
8	https://cdi	Kylian MbappÃ©	http://sofi	France	ST LW RW	K. MbappÃ©	21	90	95		231747	5'10"	161lbs	Right	91	ST	5	01-Jul-18	N/A	â‚¬105.5M	â‚¬160K	â‚¬203.1M	408
9	https://cdi	Alisson Ramses B	http://sofi	Brazil	GK	Alisson	27	90	91		212831	6'3"	201lbs	Right	90	GK	1	19-Jul-18	N/A	â‚¬62.5M	â‚¬160K	â‚¬120.3M	114
10	https://cdi	Mohamed Salah	http://sofi	Egypt	RW	M. Salah	28	90	90		209331	5'9"	157lbs	Left	90	RW	0	01-Jul-17	N/A	â‚¬78M	â‚¬250K	â‚¬144.3M	392
11	https://cdi	Sadio ManÃ©	http://sofi	Senegal	LW	S. ManÃ©	28	90	90		208722	5'9"	152lbs	Right	90	LW	0	01-Jul-16	N/A	â‚¬78M	â‚¬250K	â‚¬144.3M	410
12	https://cdi	Virgil van Dijk	http://sofi	Netherlan	CB	V. van Dijk	28	90	91		203376	6'4"	203lbs	Right	90	CB	1	01-Jan-18	N/A	â‚¬75.5M	â‚¬210K	â‚¬145.3M	316
13	https://cdi	Marc-AndrÃ© te	http://sofi	Germany	GK	M. ter Stegen	28	90	93		192448	6'2"	187lbs	Right	90	GK	3	01-Jul-14	N/A	â‚¬69.5M	â‚¬260K	â‚¬147.7M	118
14	https://cdi	Carlos Henrique	http://sofi	Brazil	CDM	Casemiro	28	89	89		200145	6'1"	185lbs	Right	89	CDM	0	11-Jul-13	N/A	â‚¬59.5M	â‚¬310K	â‚¬122M	349
15	https://cdi	Thibaut Courtois	http://sofi	Belgium	GK	T. Courtois	28	89	90		192119	6'6"	212lbs	Left	89	GK	1	09-Aug-18	N/A	â‚¬56M	â‚¬250K	â‚¬119M	86
16	https://cdi	Manuel Neuer	http://sofi	Germany	GK	M. Neuer	34	89	89		167495	6'4"	203lbs	Right	89	GK	0	01-Jul-11	N/A	â‚¬29M	â‚¬125K	â‚¬47.9M	119
17	https://cdi	Karim Benzema	http://sofi	France	CF ST	K. Benzema	32	89	89		165153	6'1"	179lbs	Right	89	CF	0	09-Jul-09	N/A	â‚¬53M	â‚¬350K	â‚¬108.7M	426
18	https://cdi	Sergio Ramos Gai	http://sofi	Spain	CB	Sergio Ramos	34	89	89		155862	6'0"	181lbs	Right	89	CB	0	01-Aug-05	N/A	â‚¬24.5M	â‚¬300K	â‚¬50.2M	374

Dataset view after cleaning

ID_proper	Full_name	Name	Age	Nationality	Position_1	Position_2	Position_3	Overall Rating	Potential Rating	Team_name	Contract
158023	Lionel Messi	L. Messi	33	Argentina	RW	ST	CF	93%	93%	FC Barcelona	2004 - 2021
020801	C. Ronaldo dos Santos Aveiro	Cristiano Ronaldo	35	Portugal	ST	LW		92%	92%	Juventus	2018 - 2022
200389	Jan Oblak	J. Oblak	27	Slovenia	GK			91%	93%	Atlético Madrid	2014 - 2023
192985	Kevin De Bruyne	K. De Bruyne	29	Belgium	CAM	CM		91%	91%	Manchester City	2015 - 2023
190871	Neymar da Silva Santos Jr.	Neymar Jr	28	Brazil	LW	CAM		91%	91%	Paris Saint-Germain	2017 - 2022
188545	Robert Lewandowski	R. Lewandowski	31	Poland	ST			91%	91%	FC Bayern München	2014 - 2023
231747	Kylian Mbappé	K. Mbappé	21	France	ST	LW	RW	90%	95%	Paris Saint-Germain	2018 - 2022
212831	Alisson Ramses Becker	Alisson	27	Brazil	GK			90%	91%	Liverpool	2018 - 2024
209331	Mohamed Salah	M. Salah	28	Egypt	RW			90%	90%	Liverpool	2017 - 2023
208722	Sadio Mané	S. Mané	28	Senegal	LW			90%	90%	Liverpool	2016 - 2023
203376	Virgil van Dijk	V. van Dijk	28	Netherlands	CB			90%	91%	Liverpool	2018 - 2023
192448	Marc-André ter Stegen	M. ter Stegen	28	Germany	GK			90%	93%	FC Barcelona	2014 - 2022
200145	Carlos Henrique Venancio Casimiro	Casemiro	28	Brazil	CDM			89%	89%	Real Madrid	2013 - 2023
192119	Thibaut Courtois	T. Courtois	28	Belgium	GK			89%	90%	Real Madrid	2018 - 2024
167495	Manuel Neuer	M. Neuer	34	Germany	GK			89%	89%	FC Bayern München	2011 - 2023
165153	Karim Benzema	K. Benzema	32	France	CF	ST		89%	89%	Real Madrid	2009 - 2022
155862	Sergio Ramos García	Sergio Ramos	34	Spain	CB			89%	89%	Real Madrid	2005 - 2021
153079	Sergio Agüero	S. Agüero	32	Argentina	ST			89%	89%	Manchester City	2011 - 2021
215914	N'Golo Kanté	N. Kanté	29	France	CDM	CM		88%	88%	Chelsea	2016 - 2023
212622	Joshua Kimmich	J. Kimmich	25	Germany	CDM	RB		88%	90%	FC Bayern München	2015 - 2023
211110	Paulo Dybala	P. Dybala	26	Argentina	CF	CAM		88%	89%	Juventus	2015 - 2022
210257	Ederson Santana de Moraes	Ederson	26	Brazil	GK			88%	91%	Manchester City	2017 - 2024

The dataset ready to be used for further analysis.