



DATA CLEANING USING POWER QUERY

FIFA21 PLAYERS DATASET

Introduction

The Data challenge space organized a data cleaning challenge for data enthusiasts to flex their data cleaning skills and I decided to participate in it.

Data cleaning is a very important stage every data must go through before carrying out analysis as the state of the data will influence the insights generated.

My guiding objective is to ensure that all the columns have the appropriate data type. To achieve this objective, the data will need to be studied and understood.

About the Dataset

The dataset provided by the data challenge space is the FIFA21 data and it was sourced from Kaggle. The data is very messy and it contains details of FIFA21 players and their performance ratings. The dataset has 18979 rows and 77 columns.

From observation, the names columns (Name & LongName), club, value, wage, release clause, have special characters. Converting the file origin to Unicode (UTF-8) before loading the data in the power query will change those affected columns to the right format.

Data Cleaning Process

The data was imported into Microsoft's Power Query Editor, file origin changed to Unicode (UTF-8) and data loaded for transformation.

File Origin		Delimiter
65001: Unicode (UTF-8)		Comma
ID	Name	LongName
158023	L. Messi	Lionel Messi
20801	Cristiano Ronaldo	C. Ronaldo dos Santos Aveiro
200389	J. Oblak	Jan Oblak
192985	K. De Bruyne	Kevin De Bruyne
190871	Neymar Jr	Neymar da Silva Santos Jr.
188545	R. Lewandowski	Robert Lewandowski
209331	M. Salah	Mohamed Salah
212831	Alisson	Alisson Ramses Becker
231747	K. Mbappé	Kylian Mbappé
192448	M. ter Stegen	Marc-André ter Stegen
203376	V. van Dijk	Virgil van Dijk
208722	S. Mané	Sadio Mané
200145	Casemiro	Carlos Henrique Venancio Casimiro

ID, Names, OVA, BOV, POT

- The 'ID' column has players unique number. No duplicates were found. The 'ID' column was renamed **Player's ID**
- The 'Name', 'LongName' and 'PlayerUrl' columns contain names of the players. The 'Name' and 'PlayerUrl' columns were removed and 'LongName' column was renamed **Player's Name**
- The 'PhotoUrl' column which contain link to view the image of the players was removed from the dataset.
- The 'OVA' 'BOV' and 'POT' columns which are the players overall analysis, best overall analysis and players potential rating are in whole numbers and should be converted to percentage data type. Custom columns for 'Player's Overall Analysis' 'Best Overall Analysis' and 'Player's Potential Rating' were created, divided by 100 and converted to the percentage data type.

Custom Column

Add a column that is computed from the other columns.

New column name

Player's Overall Analysis

Custom column formula ⓘ

= [#""IOVA"]/100

[Learn about Power Query formulas](#)

✓ No syntax errors have been detected.

OK

Cancel

Available columns

Player's ID

Player's Name

Nationality

Age

IOVA

POT

Club

<< Insert

% Player's Overall Analysis	% Player's Potential Rating
93.00%	93.00%
92.00%	92.00%
91.00%	93.00%
91.00%	91.00%

Contract Column

- The 'Contract' column contains data in different formats which needs to be cleaned for consistency in year format. For the cleaning process, the column was split using space as the delimiter. The columns with the contract start year and end year were kept and renamed. The entries in the contract column which had no start and end year were replaced with 'null' and these represented 1250 players.

- ☒ 2020 ~ 2024
- ☒ 2020 ~ 2025
- ☒ 2020 ~ 2026
- ☒ 2020 ~ 2027
- ☒ Dec 31, 2020 On Loan
- ☒ Free
- ☒ Jan 30, 2021 On Loan
- ☒ Jun 30, 2021 On Loan

1 ² ₃ Contract Start Year	1 ² ₃ Contract End Year
null	null
null	null
null	null

1 ² ₃ Contract Start Year	1 ² ₃ Contract End Year
2004	2021
2018	2022
2014	2023

Height Column

- The 'Height' column has values in centimeters and feet/inches. To ensure consistency, the height column is being changed to cm.
- Two Conditional columns are created to convert feet to inches (If "cm" then 1 else 12) and inches to cm (if "cm" then 1 else 0.3937).
- The height column is further split using "" as delimiter to separate the feet from inches and this forms Height.1 and Height.2 columns
- In order to multiply the feet in Height.1 with 12, Height.1 column is converted to whole number. After the multiplication, the result is added to the existing inches (Height.2) to get the total inches
- The total inches is multiplied with 1 and 0.3937 which converts the result to cm

1 ² ₃ Height.1	ABC 123 Height_Multiply by 12	1.2 Multiplication
6	12	72
6	12	72
6	12	72
5	12	60

1.2 Multiplication	1 ² ₃ Height.2	1.2 Addition
60	10	70
60	9	69
60	7	67
60	10	70

1.2 Addition	ABC 123 Height_Multiply by 0.3937	1.2 Multiplication.1
74	0.3937	29.1338
75	0.3937	29.5275
77	0.3937	30.3149

Weight Column

- The 'Weight' column has values in kilograms (kg) and pounds (lbs). To ensure consistency, the weight column is being changed to **kg**.
- A Conditional column is created to convert lbs to kg (If "kg" then 1 else 2.205).
- In order to divide the weight with 1 & 2.205 to convert weight to kg, the weight column is converted to whole number before division.

1 ² ₃ Weight	ABC 123 Weight_divide by 2.205	1 ² ₃ Weight (kg)
183	2.205	83
179	2.205	81
183	2.205	83

Loan Date End

- The 'Loan Date End' column is 95% empty and 5% valid. To solve this, the empty cells are replaced with "null"

Loan Date End		
1013 (5%) Valid	0 (0%) Error	17966 (95%) Empty
Remove Empty		

<i>null</i>
<i>null</i>
<i>null</i>
<i>null</i>
<i>null</i>

Value, Wage, Release Clause

- The 'Value', 'Wage' and 'Release Clause' columns have inconsistent data type with the Euro sign "€" and suffixes "m" for million, "k" for thousand

Conditional columns were created and the currencies were converted to whole numbers.

A ^B _C Value	A ^B _C Wage	A ^B _C Release Clause
€103.5M	€560K	€138.4M
€63M	€220K	€75.9M
€120M	€125K	€159.4M
€129M	€370K	€161M

1 ² ₃ Value (Euro)	1 ² ₃ Wage (Euro)	1 ² ₃ Release Clause (Euro)
103500000	560000	138400000
63000000	220000	75900000
120000000	125000	159400000
129000000	370000	161000000

W/F, SM, IR Columns

The columns 'W/F', 'SM' and 'IR' contain players ratings on a scale of 1 to 5 with a special character and in a text format.

To clean these columns, the special character was replaced, data type converted to whole number and columns renamed in full 'Weak Foot Rating', 'Skill Moves Rating' and 'Injury Rating'

1 ² ₃ Weak Foot Rating	1 ² ₃ Skill Moves Rating	1 ² ₃ Injury Rating
4	4	5
4	5	5
3	1	3
5	4	4

A/W, D/W, PAC, SHO, PAS, Hits

- The column 'A/W' was renamed in full 'Attacking Work Rate'
- The column "D/W" was renamed in full 'Defensive Work Rate'
- The column "PAC" was renamed in full 'Pace'
- The column "SHO" was renamed in full 'Shooting Attribute'
- The column "PAS" was renamed in full 'Pass Accuracy'
- The 'Hits' column has inconsistent data type in text format. It has entries with "K" suffix and numbers in whole & decimals but in text format. Conditional column was created with output of 1000 for data entries with "K" in the 'Hits' Column and 1 if otherwise. "K" in the Hits column is replaced and data type is converted to integer before multiplication.

123 Hits
771
562
150
207

	A	B	C	D	E	F	G	H	I	J	K
1	Player's ID	Player's Name	Nationality	Age	Club	Contract Start Year	Contract End Year	Positions	Height (cm)	Weight (kg)	Preferred Foot
2	158023	Lionel Messi	Argentina	33	FC Barcelona	2004	2021	RW, ST, CF	170	72	Left
3	20801	C. Ronaldo dos Santos Aveiro	Portugal	35	Juventus	2018	2022	ST, LW	187	83	Right
4	200389	Jan Oblak	Slovenia	27	Atlético Madrid	2014	2023	GK	188	87	Right
5	192985	Kevin De Bruyne	Belgium	29	Manchester City	2015	2023	CAM, CM	181	70	Right
6	190871	Neymar da Silva Santos Jr.	Brazil	28	Paris Saint-Germain	2017	2022	LW, CAM	175	68	Right
7	188545	Robert Lewandowski	Poland	31	FC Bayern München	2014	2023	ST	184	80	Right
8	209331	Mohamed Salah	Egypt	28	Liverpool	2017	2023	RW	175	71	Left
9	212831	Alisson Ramses Becker	Brazil	27	Liverpool	2018	2024	GK	191	91	Right
10	231747	Kylian Mbappé	France	21	Paris Saint-Germain	2018	2022	ST, LW, RW	178	73	Right
11	192448	Marc-André ter Stegen	Germany	28	FC Barcelona	2014	2022	GK	187	85	Right
12	203376	Virgil van Dijk	Netherlands	28	Liverpool	2018	2023	CB	193	92	Right
13	208722	Sadio Mané	Senegal	28	Liverpool	2016	2023	LW	175	69	Right
14	200145	Carlos Henrique Venancio Casimiro	Brazil	28	Real Madrid	2013	2023	CDM	185	84	Right
15	192119	Thibaut Courtois	Belgium	28	Real Madrid	2018	2024	GK	199	96	Left
16	167495	Manuel Neuer	Germany	34	FC Bayern München	2011	2023	GK	193	92	Right
17	165153	Karim Benzema	France	32	Real Madrid	2009	2022	CF, ST	185	81	Right
18	155862	Sergio Ramos García	Spain	34	Real Madrid	2005	2021	CB	184	82	Right
19	153079	Sergio Agüero	Argentina	32	Manchester City	2011	2021	ST	173	70	Right
20	202652	Raheem Sterling	England	25	Manchester City	2015	2023	LW, RW	170	69	Right
21	215814	N'Golo Kanté	France	29	Chelsea	2016	2023	CDM, CM	169	70	Right

Conclusion

At the end of the cleaning process, I had 75 columns and 18979 rows ready for analysis.

The cleaning process was a bit challenging but also very refreshing to see the messy dataset transformed into a clean state ready for use.

Thank You.