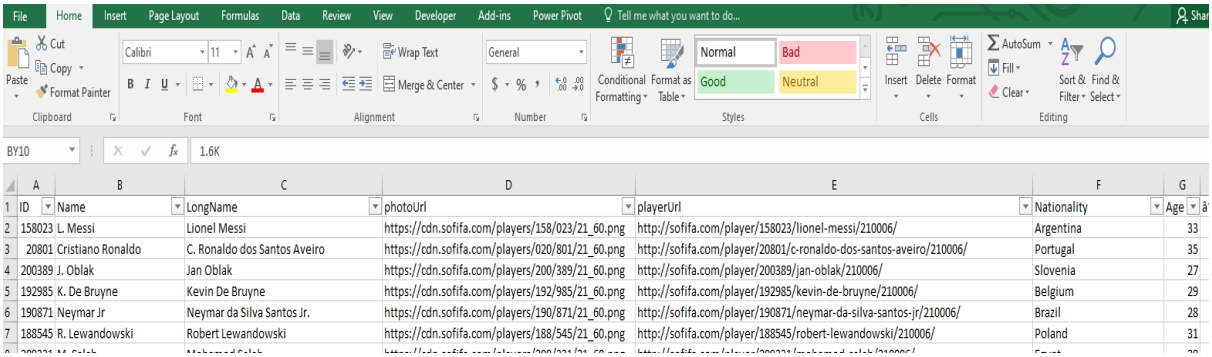
Data Wrangling process on FIFA21 raw data in Excel

Introduction

I stumbled upon a *#datacleaningchallenge* on twitter which was organized recently in the data tech space to give newbies and intermediate learners the chance to work on a data cleaning project. I created this data cleaning project utilizing the FIFA 21 dataset and Microsoft Excel tools and the goal of the data cleaning process was to ensure that the data was ready for analysis and that any errors or inconsistencies were corrected.

Data Description

The FIFA21 dataset in its messy form was gotten from kaggle and contains 18980 rows and 76 columns, which include information such as the player name, playerUrl, LongName, photoUrl, Nationality, BOV, club, Age, POT, OVA, Contract, Positions, Height, Weight, Preferred foot, Best Position etc. The dataset was in a structured format with each player's information occupying a row in the dataset. The data file also includes the data dictionary



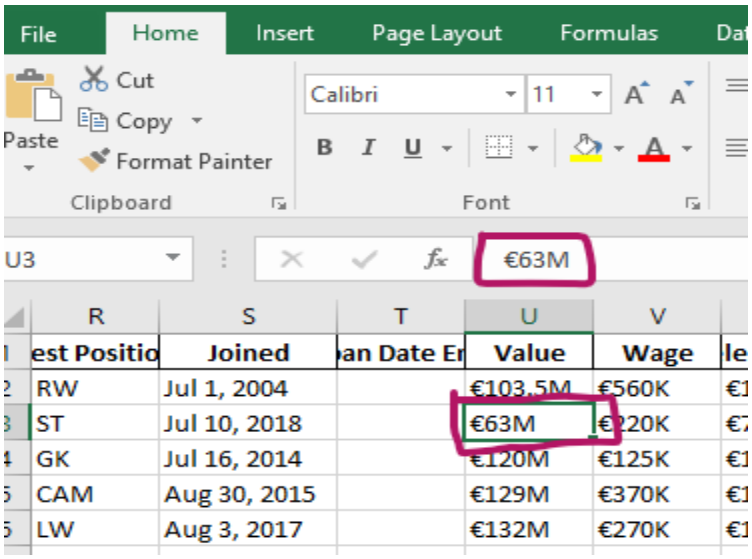
ID	Name	LongName	photoUrl	playerUrl	Nationality	Age
158023	L. Messi	Lionel Messi	https://cdn.sofifa.com/players/158/023/21_60.png	http://sofifa.com/player/158023/lionel-messi/210006/	Argentina	33
20801	Cristiano Ronaldo	C. Ronaldo dos Santos Aveiro	https://cdn.sofifa.com/players/020/801/21_60.png	http://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/210006/	Portugal	35
200389	J. Oblak	Jan Oblak	https://cdn.sofifa.com/players/200/389/21_60.png	http://sofifa.com/player/200389/jan-oblak/210006/	Slovenia	27
192985	K. De Bruyne	Kevin De Bruyne	https://cdn.sofifa.com/players/192/985/21_60.png	http://sofifa.com/player/192985/kevin-de-bruyne/210006/	Belgium	29
190871	Neymar Jr	Neymar da Silva Santos Jr.	https://cdn.sofifa.com/players/190/871/21_60.png	http://sofifa.com/player/190871/neymar-da-silva-santos-jr/210006/	Brazil	28
188545	R. Lewandowski	Robert Lewandowski	https://cdn.sofifa.com/players/188/545/21_60.png	http://sofifa.com/player/188545/robert-lewandowski/210006/	Poland	31

Club	Contract	Positions	Height	Weight	Preferred Foot	BOV	Best Position	Joined	Loan Date End	Value	Wage	Release Clause
	2004 ~ 2021	RW, ST, CF	170cm	72kg	Left	93	RW	1-Jul-04		â‚¬103.5M	â‚¬560K	â‚¬138.4M
	2018 ~ 2022	ST, LW	187cm	83kg	Right	92	ST	10-Jul-18		â‚¬63M	â‚¬220K	â‚¬75.9M
	2014 ~ 2023	GK	188cm	87kg	Right	91	GK	16-Jul-14		â‚¬120M	â‚¬125K	â‚¬159.4M
	2015 ~ 2023	CAM, CM	181cm	70kg	Right	91	CAM	30-Aug-15		â‚¬129M	â‚¬370K	â‚¬161M
	2017 ~ 2022	LW, CAM	175cm	68kg	Right	91	LW	3-Aug-17		â‚¬132M	â‚¬270K	â‚¬166.5M

The first cleaning step taken was to look for duplicate values, but no duplicate value was found. After exploring the dataset, the following issues were found in the dataset:

1. **Error in spelling**: The Name, LongName and Nationality column were found to be encoded in UTF – 8.

2. **Incorrect formatting**: Columns like Wage, Height, Weight, Value, Release Clause, Hits contain incorrect format. E.g. In the wage column, the first row represented as €560K instead of 560,000



3. **Inconsistent Formatting:** The height column contains data in cm and Feet/inches and the weight has data in kg and lbs.

M	N	O
Height	Weight	Preferr
165cm	60kg	Right
163cm	59kg	Right
163cm	61kg	Right
163cm	70kg	Right
165cm	58kg	Right
165cm	62kg	Left
6'2"	183lbs	Right
164cm	60kg	Right
5'11"	172lbs	Right
6'1"	176lbs	Right
6'0"	185lbs	Right
6'1"	179lbs	Right
5'11"	170lbs	Left
6'2"	196lbs	Right
6'0"	172lbs	Left

Data Cleaning and Transformation

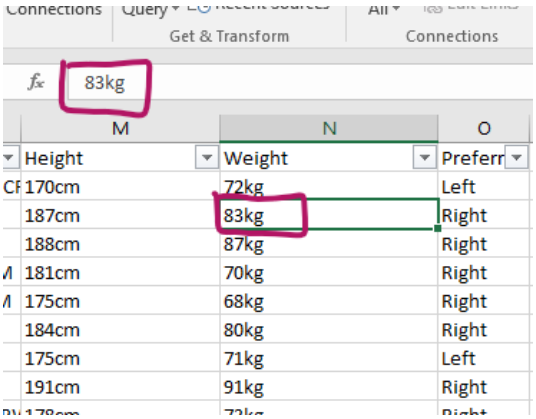
The data was loaded first on python to decode the UTF – 8 into ANSI which was then saved into an "xlsx" format. In doing this, the Name, LongName and Nationality issues ware solved.

The data was then loaded into Excel and the following transformation were made.

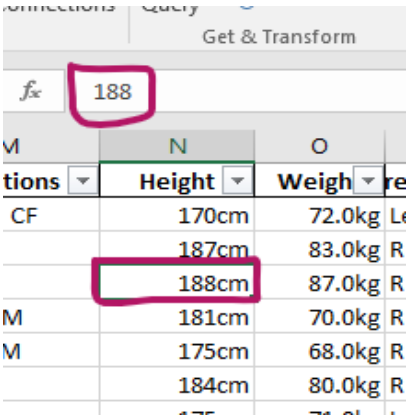
1. **Height column**
   - Replaced the 'cm' and (") character with an empty character
   - Split the column with Text to Columns tool with ' as the delimiter
   - Performed a calculation that returns the 1st column if the second column is empty and converts to "cm" if the 2nd column has a value.
   - Formatted the column into numeric with cm as unit

2. **Weight columns**
   - Replaced the “kg” and “lbs” character with “-kg” and “-lbs” respectively
   - Split the column with Text to Columns tool with - as the delimiter
   - Performed a calculation that returns the 1st column if the second column contains “kg” and converts to “kg” if the 2nd column contains “lbs”.
   - Formatted the column into numeric with kg as unit

Before (Height & Weight Column)

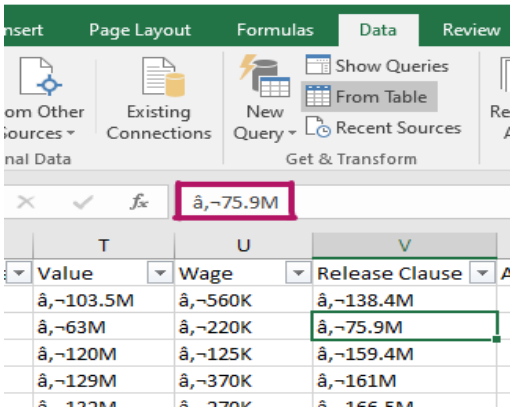


After

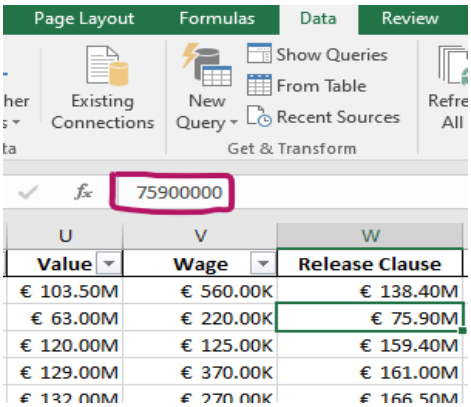


3. **Value column**
   - Replaced the ‘€’ character with an empty character
   - Created two columns to separate the numbers from text using the LEFT and RIGHT function. The LEFT function gets the number and the RIGHT gets the text (M or K)
   - Performed a calculation that multiplies the number with 10^6 or 10^3 if the text is M or K
   - Formatted the result with ‘€’ currency and millions and thousand suffix (e.g. 1000000 >> 1M)
4. **Release Clause column**
   - Repeated the steps for Value column
5. **Wage column**
   - Replaced the ‘€’ character with an empty character and the ‘K’ character with ‘–k’
   - Split the column with Text to Columns tool with - as the delimiter
   - Performed a calculation multiply the 1st column with 10^3 if the second column contains “K” and return the 1st column if the second column is empty.
   - Formatted the result with ‘€’ currency and thousand suffix (e.g. 1000

Before (Value, Wage & Release Clause)



After



6. For **W/F**, **SM**, **IR** columns, I replaced the ★ with an empty character and changed the data type to numeric
7. Converted the **Player URL** and **Photo URL** columns to hyperlink format from text using the HYPERLINK function.

Before

E
playerUrl
http://sofifa.com/player/158023/lionel-messi/210006/
http://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/210006/
http://sofifa.com/player/200389/jan-oblak/210006/
http://sofifa.com/player/192985/kevin-de-bruyne/210006/
http://sofifa.com/player/190871/neymar-da-silva-santos-jr/210006/
http://sofifa.com/player/188545/robert-lewandowski/210006/

After

F
playerUrl
http://sofifa.com/player/158023/lionel-messi/210006/
http://sofifa.com/player/20801/c-ronaldo-dos-santos-aveiro/210006/
http://sofifa.com/player/200389/jan-oblak/210006/
http://sofifa.com/player/192985/kevin-de-bruyne/210006/
http://sofifa.com/player/190871/neymar-da-silva-santos-jr/210006/
http://sofifa.com/player/188545/robert-lewandowski/210006/
http://sofifa.com/player/209331/mohamed-salah/210006/

8. Hits column
   - Replaced the 'K' character with '–k'
   - Split the column with Text to Columns tool with - as the delimiter
   - Performed a calculation multiply the 1st column with 10^3 if the second column contains "K" and return the 1st column if the second column is empty.
   - Formatted the result with thousand suffix (e.g. 1000 >> 1K)

Conclusion and Recommendation

The data cleaning process was successful in ensuring that the dataset was ready for analysis. The process resulted in a more accurate and consistent dataset, which will improve the quality of analysis

Ultimately, I have been able to improve my data cleaning abilities and learn as I go by taking on this project. Along the way I picked up new ideas and techniques, but I didn't hesitate to use what I already knew to meet the task.

It is advised that in addition to conducting research, one should take time to study the data dictionary so that he/she can comprehend the content of the dataset, this will assist in managing data quality, consistency and security for compliance audit, it also reduce the likelihood of losing crucial data and performing ineffective analyses as a result of misunderstanding and carelessness with regard to the terms and entries in the data