**FIFA 21 CLEANING DATA**



**OBJECTIVE**

This documentation seeks to highlight the data cleaning process for this challenge, which was obtained from [Kaggle](https://www.kaggle.com/datasets/yagunnersya/fifa-21-messy-raw-dataset-for-cleaning-exploring). The data cleaning process was completed using Excel. The purpose of this project was to clean and prepare the dataset for visualization.

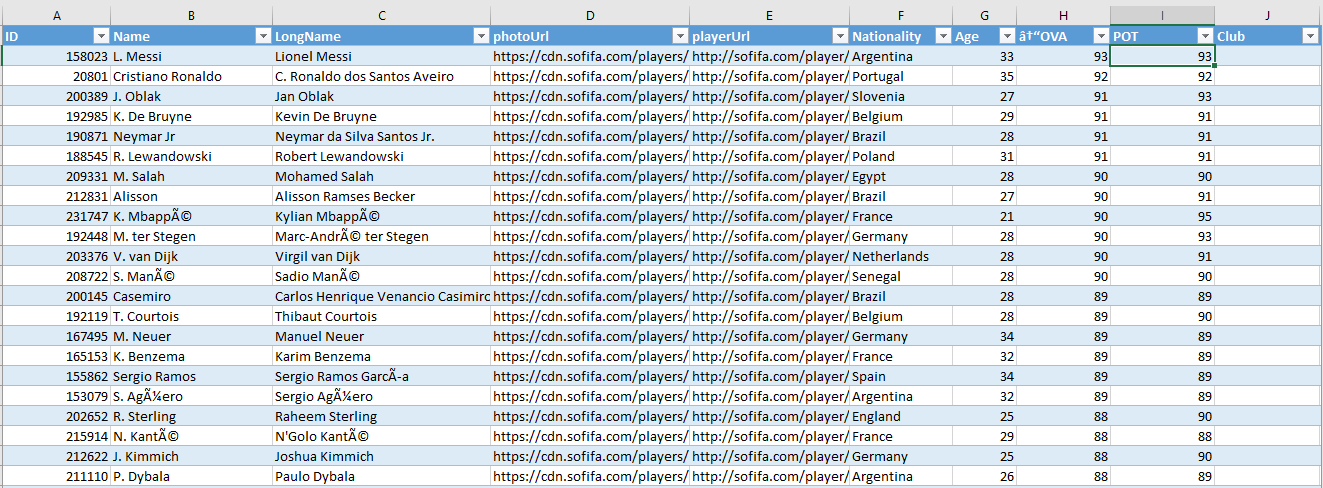
**DATA DESCRIPTION**

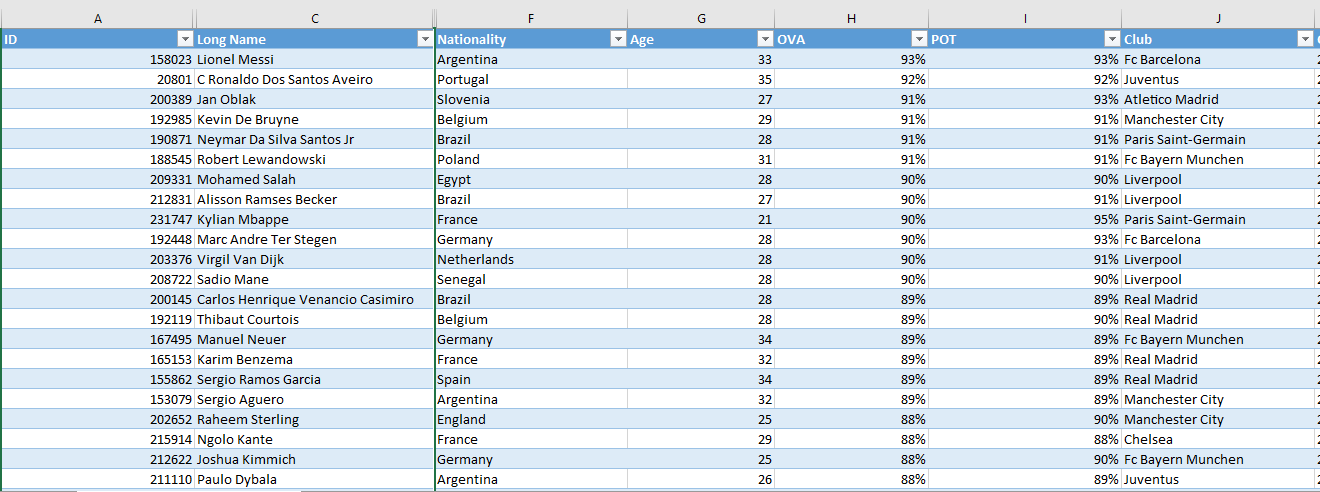
The dataset contains 18,980 rows and 77 columns. The dataset consists of different data types and contains player attributes, their clubs, nationality, how much they are worth and how much they earn.

**DATA CLEANING STEPS**

* I made a copy of the data first, so I do not lose the original data.
* I then changed the ID column from whole number to text, because we would not make any calculations from the ID column.
* Used wrap text on the Club column to remove the extra spaces.
* Used text to column to obtain the correct spelling of the names from the playerurl column.
* Removed special characters from the OVA header.

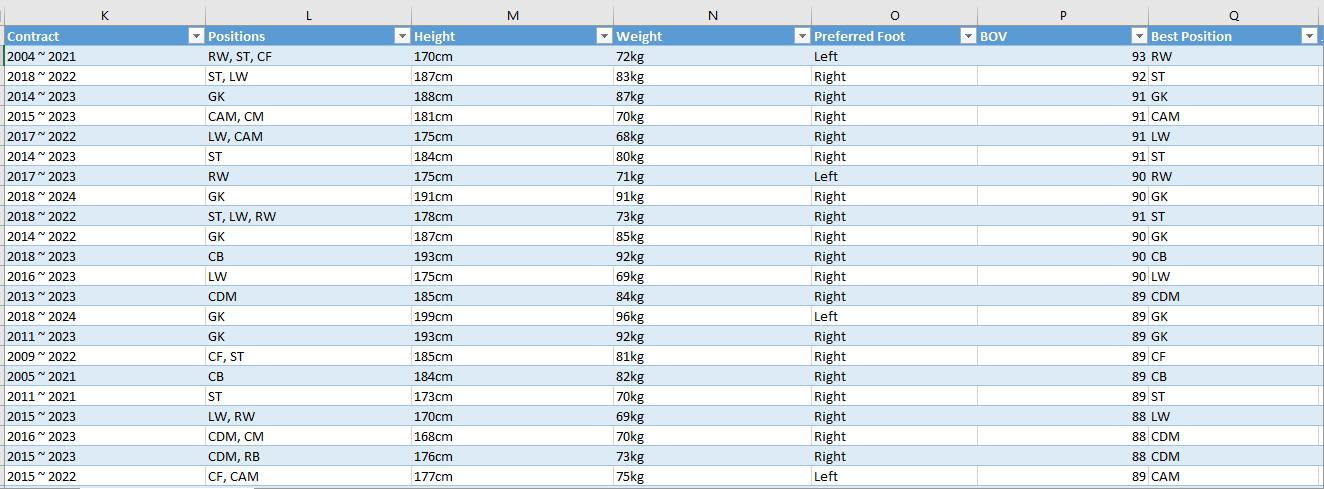
Below is the dirty and cleaned data:

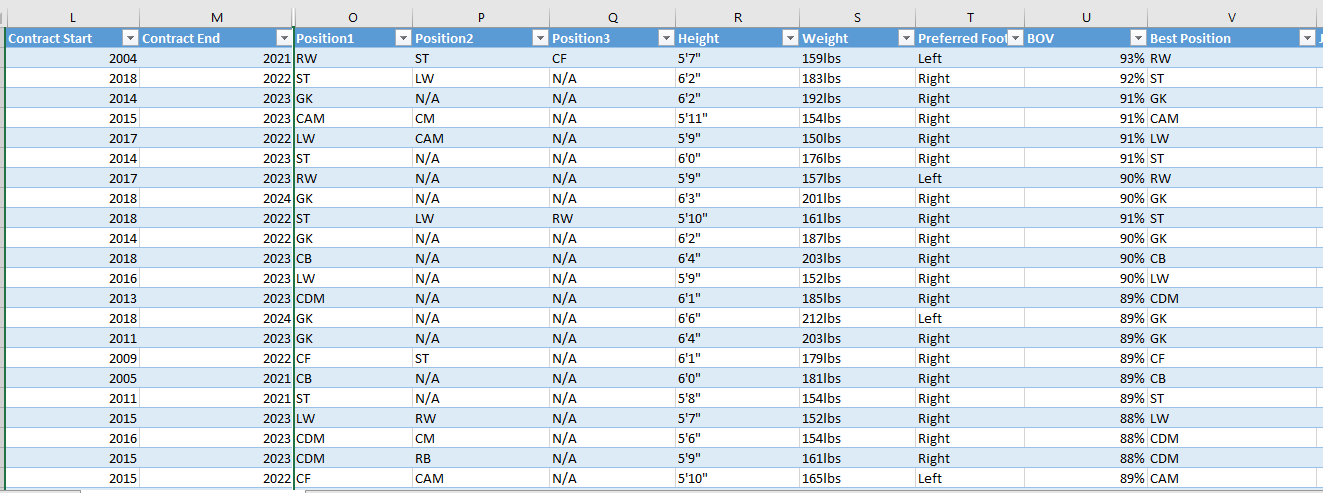




* For the contract column, it was split into the start and end date
* The positions were also split into three columns, using text-to-column, to account for players who could play more than one position.
* I changed the numbers stored as texts in the weight column to numbers using the paste special format, and added the weight unit (lbs) at the end of the numbers using the ref formula (=cell number&”lbs”)
* For the height column, I converted all numbers stored as text in cm to numbers, using the special paste format. I then converted to feet and converted it to feet per inches, using the formula [=INT(CELLREF)+(12\*MOD(CELLREF,1)>=11.5)&”’”&IF(12\*MOD(CELLREF,1)>=11.5,0,ROUND(12\*MOD(CELLREF,1),0))&””””]. I then copied and pasted the original numbers in feet per inches to the new column.
* Converted the BOV column into a percentage.

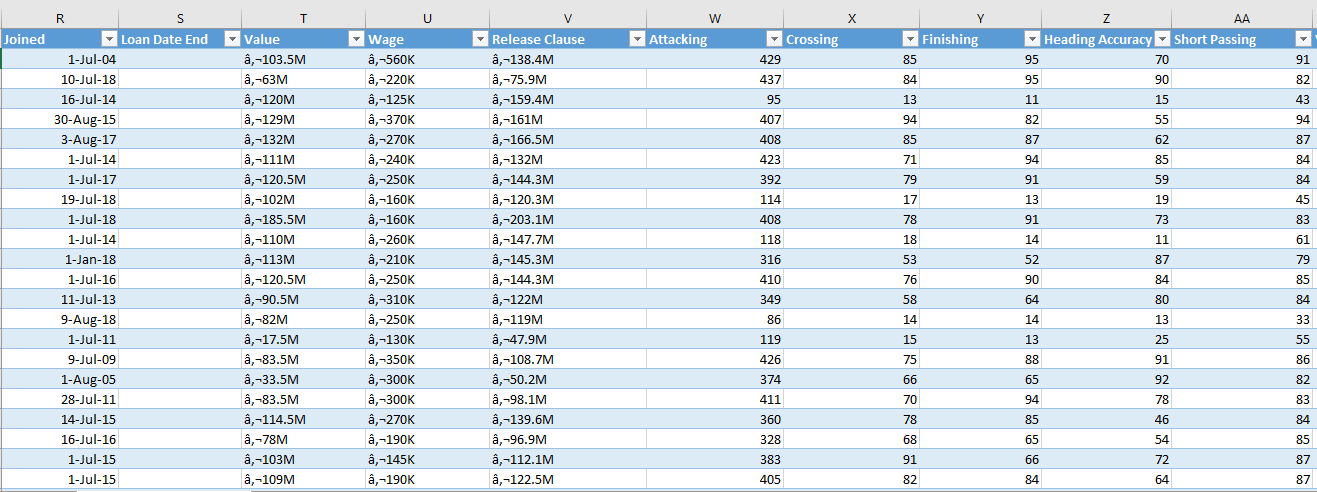
Below is the dirty and cleaned data:

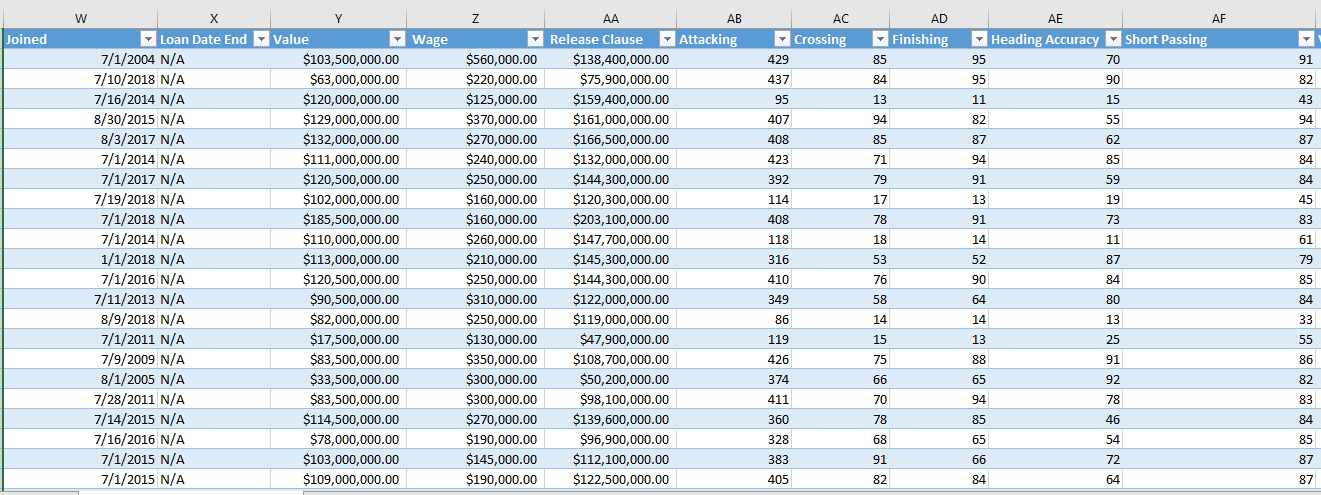




* I changed the Joined Date column to a short date and created a custom date with format MM-DD-YYYY.
* I converted numbers stored as text in the value column to numbers, by first replacing M, with blank/space, and converted by using the paste special and multiplied by 1,000,000. I did the same for the K, by using find and replace and replaced numbers with K to 000, hence, converting K to thousands.
* I applied the same format above for the wage column
* I applied the same format above for the release clause column. It is important to note that when using the format in line 3, one must convert the text column to numbers by using the paste special method, before using find and replace to change the allocated number symbol to numbers.
* I changed the special characters in the wage, release clause and value column from (-) to the zero number.

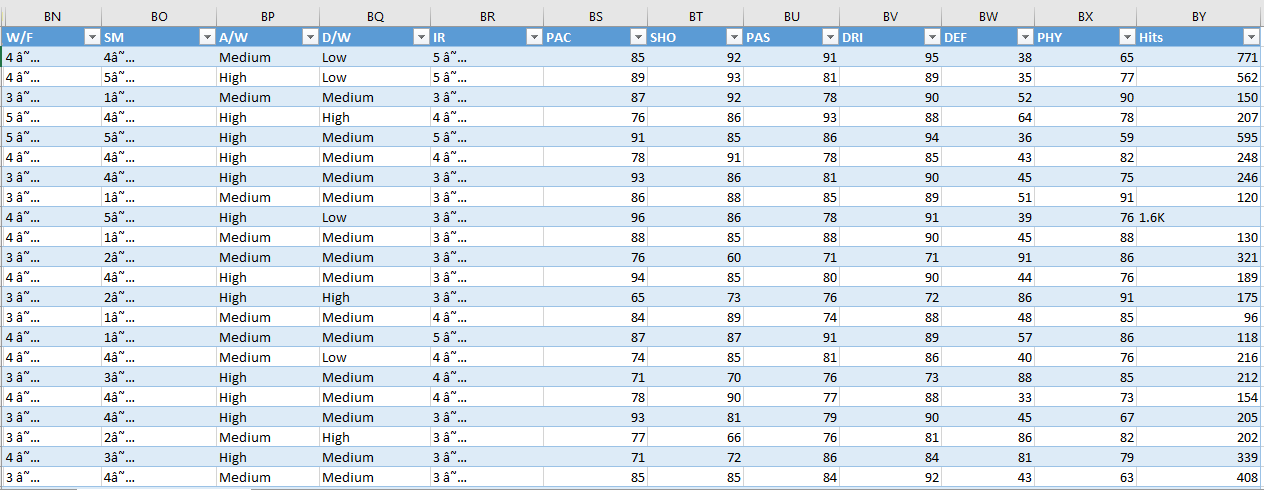
Below is the dirty and cleaned data:

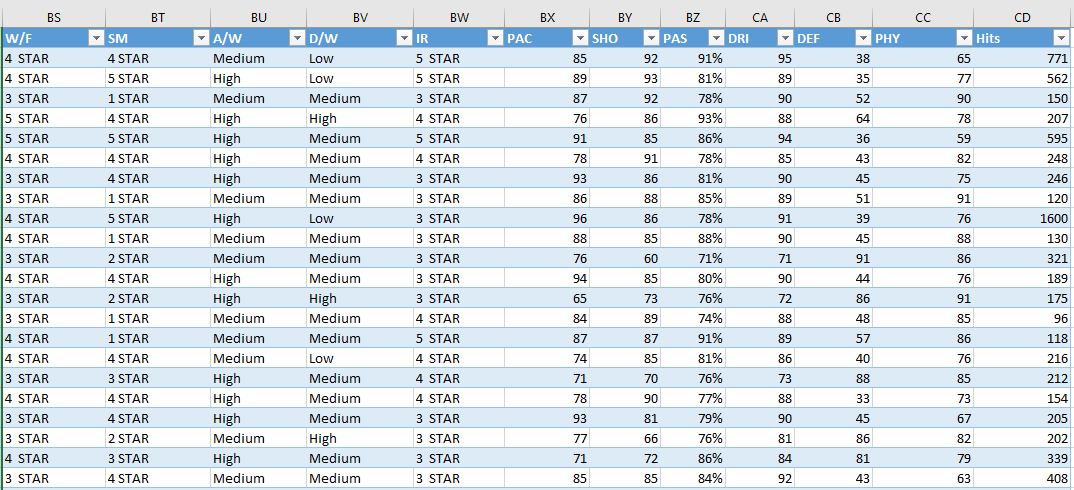




* Changed unique characters in columns W/F, SM and IR to STAR using the find and replace function.
* I used the formula [=IF(RIGHT(CD2,1)="K",LEFT(CD2,LEN(CD2)-1)\*1000,CD2\*1)] to substitute all numbers with a K attached to them to numbers in the Hits Column.

Below is the dirty and cleaned data:





**FINAL CLEANING STEPS**

* I changed the numbers in the OVA, BOV, PASS and POT column to percentages using the format (%), in the custom option of the format cells.
* Filled all blank rows with N/A

**CONCLUSION**

This project highlighted the importance of detailed data cleaning to achieve accuracy and reliability of datasets for analysis and visualization.