**TITLE:**

TURNING MESSY RAW DATA TO CLEAN DATA USING MICROSOFT EXCEL

**THE DATA SET:**

This is a real world dataset of over 18,000 records of FIFA21 players' showing players' name, age, club, photo URL, salary, etc. All in dirty form.

**DATA SOURCE:**

Kaggle https://www.kaggle.com/datasets/yagunnersya/fifa-21-messy-raw-dataset-for-cleaning-exploring

**BRIEF GUIDE ABOUT THE DATASET:** This describes the field names in the dataset

- ID: The unique identifier for the player.

- HEIGHT: The height of the player in feet and inches.

- WEIGHT: The weight of the player in pounds.

- FOOT: The preferred foot of the player.

- BOV: The best overall rating the player has achieved in their career.

- BP: The best position the player has played in their career.

- GROWTH: The difference between the potential rating and overall rating of the player.

- JOINED: The date the player joined their current team in FIFA 21.

- LOAN DATE END: The date the player's loan contract ends.

- VALUE: The market value of the player in FIFA 21.

- WAGE: The weekly wage of the player in FIFA 21.

- RELEASE CLAUSE: The release clause value of the player in FIFA 21.

- ATTACKING: The attacking attributes of the player.

- CROSSING: The crossing attribute of the player.

- FINISHING: The finishing attribute of the player.

- HEADING ACCURACY: The heading accuracy attribute of the player.

- SHORT PASSING: The short passing attribute of the player.

- VOLLEYS: The volleys attribute of the player.

- SKILL: The skill attributes of the player.

- DRIBBLING: The dribbling attribute of the player.

- CURVE: The curve attribute of the player.

- FK ACCURACY: The free kick accuracy attribute of the player.

- LONG PASSING: The long passing attribute of the player.

- BALL CONTROL: The ball control attribute of the player.

- MOVEMENT: The movement attributes of the player.

- ACCELERATION: The acceleration attribute of the player.

- SPRINT SPEED: The sprint speed attribute of the player.

- AGILITY: The agility attribute of the player.

- REACTIONS: The reactions attribute of the player.

- BALANCE: The balance attribute of the player

- POWER: The power attributes of the player.

- SHOT POWER: The shot power attribute of the player.

- JUMPING: The jumping attribute of the player.

- STAMINA: The stamina attribute of the player.

- STRENGTH: The strength attribute of the player.

- LONG SHOTS: The long shots attribute of the player.

- MENTALITY: The mentality attributes of the player.

- AGGRESSION: The aggression attribute of the player.

- INTERCEPTIONS: The interceptions attribute of the player.

- POSITIONING: The positioning attribute of the player.

- VISION: The vision attribute of the player.

- PENALTIES: The penalties attribute of the player.

- COMPOSURE: The composure attribute of the player.

- DEFENDING: The defending attributes of the player.

- MARKING: The marking attribute of the player.

- STANDING TACKLE: The standing tackle attribute of the player.

- SLIDING TACKLE: The sliding tackle attribute of the player.

- GOALKEEPING: The goalkeeping attributes of the player.

- GK DIVING: The goalkeeper diving attribute of the player.

- GK HANDLING: The goalkeeper handling attribute of the player.

- GK KICKING: The goalkeeper kicking attribute of the player.

- GK POSITIONING: The goalkeeper positioning attribute of the player.

- GK REFLEXES: This refers to the goalkeeper's ability to react and make saves quickly.

- TOTAL STATS: This refers to the overall rating of the player based on their performance in all areas of the game.

- BASE STATS: This refers to the player's rating in the six main areas of the game: Pace, Shooting, Passing, Dribbling, Defending, and Physicality.

- W/F: This refers to the player's weaker foot ability.

- SM: This refers to the player's skill moves ability.

**THE TASK:**

Completely clean the dataset and make it fit for analysis purpose.

**TOOL USED:**

Microsoft Excel

**THE LINK TO DOWNLOAD CLEANED DATA**

Git Hub link to the Clean Data

**DOCUMENTATION OF STEPS IN ACHIEVING A CLEAN DATASET:**

I first of all tested for duplicate records, using the ID column as rule. I formatted the dataset as table. While exploring the dataset to get acquainted with it and to determine which column(s) needs to be cleaned I found the following columns as dirty:

**(1) Name**: this column particularly had characters that were not consistent with the player's name. In solving this, I took cue from the player URL column which displayed the name of the player with hyphen as delimiter. I simply used the Microsoft Excel 's find and replace **(Ctrl+H)** feature to solve this.

Although, one would argue that splitting the player URL column to extract the names would have done the job. However, across the dataset, there were still presence of these characters across many other columns. So, it's only but a necessity to get follow the approach I used. As replacing one character/symbol with the appropriate alphabet affects sometimes over 1000 other records.

So this approach wasn't only changing for name, it was also affecting the long name column and every other columns where that same character appeared.

**(2) Long Name:** this column and the name column had same dirty situation. So, I did the two concurrently. Because while changing for one, it's also in most cases affecting not just the second column but across the dataset.

**(3) Growth:** In getting this I subtracted OVA from POT

**(4) Club:** this column showed empty even though data was in it. What I did was to textwrap the column. And that was enough to make it show

**(5) Contract:** this column had a character that is in appropriate to signify the range of two years. So I had to use the **=substitute()** function to replace

**(6) Values, Height, Weight, Release clause, Wage:** a quick exploration of these columns showed they all had one thing in common as it concerns the state of their dirtiness. Each of these columns had multiple and inconsistent (in context of the football domain) unit of measurement.

For example: In weight column, the units found there included: kg and lbs(lbm). Whereas it should be lbs(lbm)

In Height column, the units found there included: cm and inches. Whereas, it should be foot(ft) and inches

While for value, wage and release clause columns all had both M to represent Million and K to represent Thousand across the columns, but the inclusion of the letters made it impossible to work with. So, instead of 1M, I represented it as 1,000,000. Instead of 1k, I represented it with 1,000 all in dollars. More so, special characters existed in the three columns which I used the find and replace feature to sort out.

In solving for each of these five columns, I did the following:

(a) I copied out the field (one at a time) to another worksheet

(b) Formatted the column as a table

(c) Filtered by the alphabet (click on the filter button on the column field name > click filter > click contain > type in K if you're working on the wage column) With this, you'd only have values with K

(d) Like two columns away from the column, enter the function that would help you automatically remove the letter K. It involves combining LEFT function and LEN function example: =LEFT(A5, LEN(A5)-1). The -1 removes the K that's attached to the values

(e) Turn this values to thousands by multiplying by \*1000 and using auto fill for the other values. Copy and paste the result as values in the next column but ensure that you maintain same row number while being watchful in any case of value change.

(f) I repeated these steps for the values with M staring from filtering for cells that contain M but multiplying with 1000000. Also watching out for outliers.

(g) Now that I have different values in different columns. I Filtered again for K and deleted the values for K. The essence of this is to have empty cells where I can move the the thousand values in the other column to them

(h) I removed the filters and all values are now appearing. Showing empty cells where I can now move the values into

(I) I copied and pasted as values to the main worksheet

**I repeated the steps for all columns that had multiple units.**

While for weight column that had kg and lbs. Adding to the above steps, I converted kg to lbs using the CONVERT function. For example: =Convert(A5, "kg","lbs)

Things important to note is that: the data is not sorted. So just converting all entries to lbs would make the data lack integrity.

Best practice is to first convert all lbs data to kg =CONVERT(A19," lbs"," kg")

Now that all entries are in kg, I then converted them back to the standard lbs(lbm) and pasted the values back to where they originally were.

**(7) Loan Date End:** since not all players were on loan, I simply had to fill the empty spaces where the player wasn’t on loan with not applicable (N/A).

**(8)** I tested for data validity in all columns and watched out for outliers

**SKILLS USED:** Microsoft Excel, Problem solving skill, Research, Report documentation, Critical thinking, and eye for details.

**THE CONVENER:** [Promise Chinonso](https://www.linkedin.com/in/promisechinonso). She challenged the data community with this task. Kudos to her!