Project 1

* Data Cleaning & Transformation

In this project, we will work with a dataset from FIFA 2021 players.

The dataset is messy and our task includes cleaning the data and transforming it into a readable format.

Instructions:

1. Convert the height and weight columns to numerical forms
2. Remove the unnecessary newline characters from all columns that have them.
3. Based on the 'Joined' column, check which players have played at a club for over 10 years!
4. 'Value', 'Wage' and "Release Clause' are string columns. Convert them to numbers. For eg, "M" in the value column is Million, so multiply the row values by 1,000,000, etc.
5. Some columns have 'star' characters. Strip those columns of these stars and make the columns numerical
6. Which players are precious but still underpaid (on low wages)? (hint: scatter plot between wage and value)
7. Do the height and weight columns have the appropriate data types?
8. Can you separate the joined column into year, month, and day columns?
9. Can you clean and transform integer value, wage, and release clause columns?
10. How can you remove the newline characters from the Hits column?
11. Should you separate the Team Contract column into separate team and contract columns?

Recommended tools:

Python or Excel

Excel:

1. We have converted the Height and Weight columns to numbers.
2. We have removed the unnecessary newline characters from all columns that have them, first by selecting *ctrl+A* all cells with values and then opening the “File and Replace” dialog box, and in the Find we have typed ‘*ctrl+J’* (this represents the newline character in Excel), and in the Replace filed we did not make any changes.
3. To check which players have played at a club for over 10 years based on the “Joined” column in Excel, we have calculated the difference between the current year and the year they joined the club

A new column was added, it was called “Years at the Club”. The following formula at the column was added: =DATEDIF(R2, 2021(2021,1,1),”Y”), and after we have set the format to ‘Number’ to display the result as a number of years. . After that, we highlighted the cells with the players that have more than 1- year in one club with the help of the Conditional Formatting tool.

1. We have changed the Wage and Released Clause with proper Euro Sign instead of the sign that was coming with the imported data(default sign), and after that we have used the following formula:

=TEXT(VALUE(SUBSTITUTE(SUBSTITUTE(Z2, "K", ""), "€", "")) \* 1000, "0") & "€" for both columns respectively.

We just made a change in the formula, “M” instead of “K” and 1000000 for 1000 respectively.

1. Height and Weight are in the appropriate data file – Number;
2. We have separated the “Joined” column into a 3 separate columns Month/Day/year ( with separate formula for each column and then right-click on the column and format cells – set to number followed by OK
3. We have replaced the Blank cells with the “/” symbol;
4. The newline characters are removed from the Hits column with =TRIM(SUBSTITUTE(CG, CHAR(10), ""))

-After that we have created a new column Cleaned without “K” sign and execute another formula:

=IF(RIGHT(CF2,1)="K", LEFT(CF2, LEN(CF2)-1) \* 1000, CF2)

Final Checks:

**1. Check for Consistency and Accuracy**

* **Verify Data Types:** Ensure that each column has the correct data type. For example, dates should be in date format, numbers should be numeric, and text should be in text format.
  + In Excel: Go to the "Data" tab and use "Text to Columns" if needed to ensure correct formatting.
* **Check for Errors or Inconsistencies:**
  + Look for any cells with errors (e.g., #VALUE!, #N/A, #DIV/0!) and correct them.
  + Check for any inconsistencies, such as unexpected text in numeric columns or mismatched date formats.

**2. Handle Missing or Blank Values**

* **Identify Missing Data:** Use Excel’s filtering or conditional formatting to find any blank cells or missing values.
* **Fill or Replace Missing Values:** Decide on a strategy for handling missing values, such as filling them with a default value or using interpolation.

**3. Validate Data Ranges and Logical Checks**

* **Check Ranges:** Ensure that the values fall within expected ranges (e.g., no negative values in a column that should only have positive numbers).
* **Apply Logical Tests:** For example, dates should make sense (e.g., no dates in the future if they are historical data).

**4. Remove Duplicates**

* **Check for Duplicates:** Use Excel’s "Remove Duplicates" feature to ensure there are no duplicate rows, unless duplicates are expected in your dataset.

**5. Review Data Formatting**

* **Ensure Proper Formatting:** Ensure that the data is formatted consistently, especially if you are combining multiple data sources.
* **Remove Extra Spaces:** Use the TRIM function to remove any extra leading or trailing spaces.