INTRODUCTION

The purpose of this project is to clean a dataset of FIFA player data and prepare it for analysis. The dataset contains information on 18,207 players, including their personal details, physical attributes, performance statistics, and market values. The dataset was obtained from the Kaggle website and is provided in a CSV format. The original dataset contains 104 columns, but for the purpose of this project, we will focus on 77 columns that require cleaning.

Data Cleaning Steps:

To clean the data, we will first load all necessary libraries that could be used along the journey of cleaning the data. We will then load our datasets and name it "fifa_dirty". The next step is to get some general information about the data and check what columns needed to be cleaned up. The dataset, from the brief information given above, shows that we have 77 columns to work with. This is a large number, and our Jupyter notebook cannot give us a clear picture of every column at a glance. Because we want to familiarize ourselves with every column's data, we would divide our data along the columns to four datasets and check for missing values and know those with special characters that need to be dropped.

After checking all our columns, we will extract the columns that need cleaning into a single dataset and start working on them from column to column.

CLUB'S COLUMN:

To clean the Club column, we will import regular expressions and write a function to drop special characters in the Club column.

CONTRACT COLUMN:

To clean the Contract column, we will call our column as a string and split along ~.

HEIGHT AND WEIGHT:

To clean the Height and Weight columns, we will check for unique values to see if only 'cm' and 'kg' respectively are the only characters we want to remove or if other characters are there. To clean this, we can do two things: 1. Drop the index from the back or 2. Add the characters to our regex and drop. We will choose to do the first method.

We will then write a function to change Feet and Inches to Centimetres. We will also put the Weight column in the same process as that of Height in our cleaning. The unique() function above shows us that

we have pounds (lbs) to convert to kg. We will write a function just as we did with the Height column to do this.

JOINED COLUMN:

Next, we will write a function to change our “Joined” column to datetime.

VALUE COLUMN:

To clean the Value column, we can write a Python function to write the figures in full. Before that, we will add the Euro (€) character to our regex function above so that we can use it to drop it. We will put the sign back in the column’s name.

WAGE AND RELEASE CLAUSE COLUMNS:

We will apply the full_figure() function to our next two columns, i.e., the Wage and Release Clause columns.

“W/F,” “SM,” and “IR” COLUMNS:

The next step is to drop the star characters in the “W/F,” “SM,” and “IR” columns. We will use the lambda function to clean all of W/F, SM, IR.

HIT COLUMN:

Our dataset is almost attaining the cleaned status. The only column left is the Hits column. We will check it out and work on it. From the heatmap above, the Hits column has a few unclean data. The null value is small. Is it? Perhaps that shouldn't affect the outcome of our analysis if we want to go ahead with analysis. But before then let's check the amount of NaN, then we decide either to drop them or fill with something else.

It looks as if we have more null value in our hands than expected, let's see if we can fill with mean() or better still, ZERO. but before that, let’s check the total amount of null value.

Instead of dropping it and missed out of the information for the other columns attached to the null value, the best thing to do is fill the missing value with zero. We need to convert the K to 1000 as display in the unique(). the best is to call our full_figure() function.

LASTLY:

To complete the documentation process, we need to:

1. Merge our cleaned data to the original dataset using the ID column as the key:
2. Reset the index to the ID column and rearrange the column positions if necessary
3. Save the cleaned and merged dataset back to the directory:

By completing these steps, we have a cleaned and merged dataset that is ready for further analysis.