Scope and Project Steps

Scope

This project aims to clean and transform the Fifa23 players dataset.

Project Steps

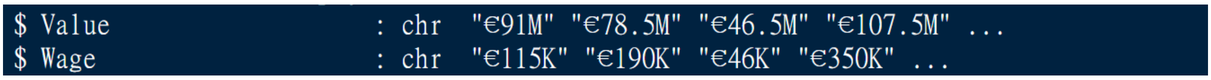
- Download the dataset from the Kaggle website then upload it into Rstudio.
- Transform the height and weight column to the appropriate data type.
- Separate the ‘Joined’ column into year, month, and day columns.
- Transform the value, wage, and release clause columns into columns of integers.
- Determine the highest-paid players for each country.

Step 1: Download and upload the Fifa23 dataset

The Fifa23 dataset was uploaded using the read.csv function. The dataset was saved as Fifa23_ds.

The str function was used to explore the data for any formatting issues. For example, the wage column was uploaded in the wrong format as a character instead of the correct number format see an image of a snippet from the str of the Fifa dataset.

‘str’ of Fifa23 dataset



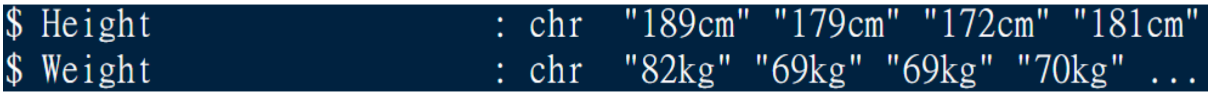
```
$ Value                    : chr  "€91M" "€78.5M" "€46.5M" "€107.5M" ...
$ Wage                     : chr  "€115K" "€190K" "€46K" "€350K" ...
```

Step 3: Transformation of the Height and Weight column

Initially, the height and weight columns were in the character format. To transform both the height and weight to the numeric format, I had to remove the characters from the numbers. Please see Image 1 for the characters to be removed (cm and kg)

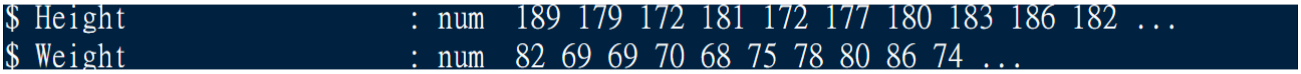
Image 1: Height and weight data presented from str function

original



```
$ Height                 : chr  "189cm" "179cm" "172cm" "181cm"
$ Weight                 : chr  "82kg" "69kg" "69kg" "70kg" ...
```

Updated



```
$ Height                    : num  189 179 172 181 172 177 180 183 186 182 ...
$ Weight                    : num  82 69 69 70 68 75 78 80 86 74 ...
```

The characters were removed using the gsub function to substitute the cm and kg for blank spaces, the blank spaces are represented by →“ ”.

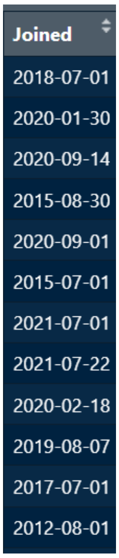
Once the characters are removed from the height and weight columns, the as.numeric function can be used to transform the height and weight columns into numeric(see updated str Image 1).

Step 3: Separate the ‘Joined’ column into year, month, and day columns

Firstly I had to transform the ‘Joined’ column from a character column into a date column using the as.date function. The %b, %d, and %Y represent the format the date will be presented in, in this case, the format is %Y = YYYY, b%=mm, and d%=dd.
I separated the dates into individual columns using the lubridate package, The lubridate function allowed me to format the ‘Joined’ column into ‘ymd’ which represents year month, and day, I was then able to pull out the year, month, and day into separate columns.

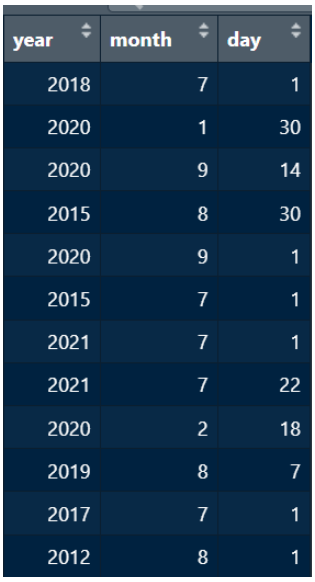
Image 2: ‘Joined’ column(the date column)

Original



Joined
2018-07-01
2020-01-30
2020-09-14
2015-08-30
2020-09-01
2015-07-01
2021-07-01
2021-07-22
2020-02-18
2019-08-07
2017-07-01
2012-08-01

Updated

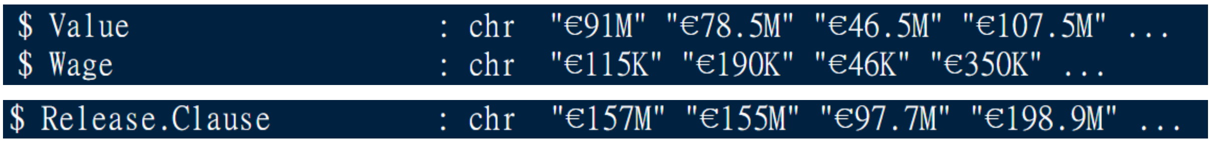


year	month	day
2018	7	1
2020	1	30
2020	9	14
2015	8	30
2020	9	1
2015	7	1
2021	7	1
2021	7	22
2020	2	18
2019	8	7
2017	7	1
2012	8	1

Step 4: Transform the value, wage, and release clause columns into columns of integers

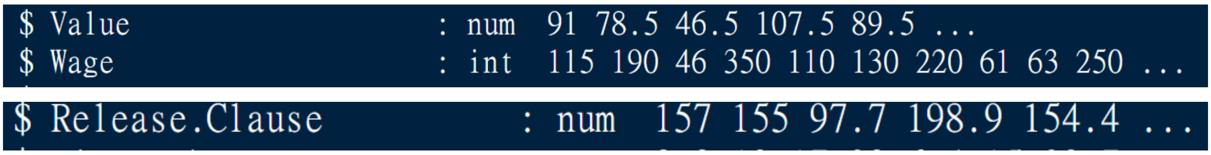
Firstly the gsub function was used to remove all string characters (K, €, and M). Then the columns were transformed to integers using the as.integer function. N/A's were introduced for the blank spaces in the value and release clause columns, I changed the N/A's in both columns to 0 by indexing 0 for n/a's I did this because I believe this is the best way to represent the empty spaces for both columns.

Original



```
$ Value              : chr  "€91M" "€78.5M" "€46.5M" "€107.5M" ...
$ Wage               : chr  "€115K" "€190K" "€46K" "€350K" ...
$ Release.Clause     : chr  "€157M" "€155M" "€97.7M" "€198.9M" ...
```

Updated



```
$ Value              : num  91 78.5 46.5 107.5 89.5 ...
$ Wage               : int  115 190 46 350 110 130 220 61 63 250 ...
$ Release.Clause     : num  157 155 97.7 198.9 154.4 ...
```

Step 5: Determine the highest-paid player by country

I created a dataset called six_figure_players, then I used the filter function from the dplyr package to filter for players for wages of 100K or more, then I grouped the players by nationality, I used the summarise function because it must be used when using a group by function, I then created a column called highest_paid_countries and set the wage to max to determine the highest paid players for each country.