FIFA Project

Introduction

This project focuses on cleaning and transforming data to ensure consistency and accuracy. The FIFA dataset provides a rich set of attributes about players but contains several inconsistencies in data formats. This includes non-standard formats in value columns, varied height and weight measurements, special characters, and mixed units that required cleaning. These inconsistencies could hinder analysis and need to be standardized to improve usability.

Therefore, the cleaning process focused on transforming the data to a standard format suitable for effective analysis.

Objective

The aim of this project was to clean and transform a raw dataset of FIFA player information to ensure consistency, accuracy, and readiness for analysis.

Dataset Summary

The dataset consisted of 18,980 rows and 75 columns, covering various attributes of FIFA players. Below is a step-by-step breakdown of the data cleaning process, including before-and-after snapshots of the dataset to illustrate the transformation process.

Data Cleaning Process

1. **Importing the Dataset**

   The dataset was imported into Excel, where initial exploratory analysis identified key inconsistencies in formats. It was then imported into Power Query within Excel for efficient transformation and data cleaning. Below is a snapshot of the dataset before cleaning.

Contract	Positions	Height	Weight	Preferred Foot	BOV	Best Position	Joined	Value	Wage	Release Clause
93 2004 ~ 2021	RW, ST, CF	170cm	72kg	Left	93	RW	01-Jul-04	â,¬103.5M	â,¬560K	â,¬138.4M
92 2018 ~ 2022	ST, LW	187cm	83kg	Right	92	ST	10-Jul-18	â,¬63M	â,¬220K	â,¬75.9M
93 2014 ~ 2023	GK	188cm	87kg	Right	91	GK	16-Jul-14	â,¬120M	â,¬125K	â,¬159.4M
91 2015 ~ 2023	CAM, CM	181cm	70kg	Right	91	CAM	30-Aug-15	â,¬129M	â,¬370K	â,¬161M
91 2017 ~ 2022	LW, CAM	175cm	68kg	Right	91	LW	03-Aug-17	â,¬132M	â,¬270K	â,¬166.5M
91 2014 ~ 2023	ST	184cm	80kg	Right	91	ST	01-Jul-14	â,¬111M	â,¬240K	â,¬132M
90 2017 ~ 2023	RW	175cm	71kg	Left	90	RW	01-Jul-17	â,¬120.5M	â,¬250K	â,¬144.3M
91 2018 ~ 2024	GK	191cm	91kg	Right	90	GK	19-Jul-18	â,¬102M	â,¬160K	â,¬120.3M
95 2018 ~ 2022	ST, LW, RW	178cm	73kg	Right	91	ST	01-Jul-18	â,¬185.5M	â,¬160K	â,¬203.1M
93 2014 ~ 2022	GK	187cm	85kg	Right	90	GK	01-Jul-14	â,¬110M	â,¬260K	â,¬147.7M
91 2018 ~ 2023	CB	193cm	92kg	Right	90	CB	01-Jan-18	â,¬113M	â,¬210K	â,¬145.3M
90 2016 ~ 2023	LW	175cm	69kg	Right	90	LW	01-Jul-16	â,¬120.5M	â,¬250K	â,¬144.3M
89 2013 ~ 2023	CDM	185cm	84kg	Right	89	CDM	11-Jul-13	â,¬90.5M	â,¬310K	â,¬122M
90 2018 ~ 2024	GK	199cm	96kg	Left	89	GK	09-Aug-18	â,¬82M	â,¬250K	â,¬119M
89 2011 ~ 2023	GK	193cm	92kg	Right	89	GK	01-Jul-11	â,¬17.5M	â,¬130K	â,¬47.9M
89 2009 ~ 2022	CF, ST	185cm	81kg	Right	89	CF	09-Jul-09	â,¬83.5M	â,¬350K	â,¬108.7M
89 2005 ~ 2021	CB	184cm	82kg	Right	89	CB	01-Aug-05	â,¬33.5M	â,¬300K	â,¬50.2M
89 2011 ~ 2021	ST	173cm	70kg	Right	89	ST	28-Jul-11	â,¬83.5M	â,¬300K	â,¬98.1M
90 2015 ~ 2023	LW, RW	170cm	69kg	Right	88	LW	14-Jul-15	â,¬114.5M	â,¬270K	â,¬139.6M

2. **Splitting Contract Dates**

- **Issue**: The "Contract" column had combined start and end dates, making it difficult to analyze contract durations.
- **Solution**: The "Contract" column was separated into "Contract Start Date" and "Contract End Date using the split by delimiter function. This enables effective analysis of players' contract periods

3. **Standardizing Columns**

- **Issue**: Columns like "Value," "Wage," and "Release Clause" included special characters and abbreviations (e.g., â,¬103.5M, â,¬560K), which were inconsistent and non-numeric.
- **Solution**: Applied the "Split by Non-Digit to Digit" function to effectively remove special characters and retain numeric values. Then used the Find and Replace function to remove abbreviations (M for million, K for thousand) into numerical values by multiplying them by the appropriate factor (1,000,000 for M, 1,000 for K) accordingly to ensure consistent data quality. Below is the after snapshot of the dataset.

Contract_Start Date	Contract_End Date	Positions	Preferred Foot	BOV	Best Position	Joined_Day	Joined_Month	Joined_Year	Value	Wage	Release Clause
2004	2021	RW, ST, CF	Left	93	RW	1	7	2004	103500000	560000	138400000
2018	2022	ST, LW	Right	92	ST	10	7	2018	63000000	220000	75900000
2014	2023	GK	Right	91	GK	16	7	2014	120000000	125000	159400000
2015	2023	CAM, CM	Right	91	CAM	30	8	2015	129000000	370000	161000000
2017	2022	LW, CAM	Right	91	LW	3	8	2017	132000000	270000	166500000
2014	2023	ST	Right	91	ST	1	7	2014	111000000	240000	132000000
2017	2023	RW	Left	90	RW	1	7	2017	120500000	250000	144300000
2018	2024	GK	Right	90	GK	19	7	2018	102000000	160000	120300000
2018	2022	ST, LW, RW	Right	91	ST	1	7	2018	185500000	160000	203100000
2014	2022	GK	Right	90	GK	1	7	2014	110000000	260000	147700000
2018	2023	CB	Right	90	CB	1	1	2018	113000000	210000	145300000
2016	2023	LW	Right	90	LW	1	7	2016	120500000	250000	144300000
2013	2023	CDM	Right	89	CDM	11	7	2013	90500000	310000	122000000
2018	2024	GK	Left	89	GK	9	8	2018	82000000	250000	119000000
2011	2023	GK	Right	89	GK	1	7	2011	17500000	130000	47900000
2009	2022	CF, ST	Right	89	CF	9	7	2009	83500000	350000	108700000
2005	2021	CB	Right	89	CB	1	8	2005	33500000		50200000
2011	2021	ST	Right	89	ST	28	7	2011	83500000	300000	98100000
2015	2023	LW, RW	Right	88	LW	14	7	2015	114500000	270000	139600000
2016	2023	CDM, CM	Right	88	CDM	16	7	2016	78000000	190000	96900000
2015	2023	CDM, RB	Right	88	CDM	1	7	2015	103000000	145000	112100000

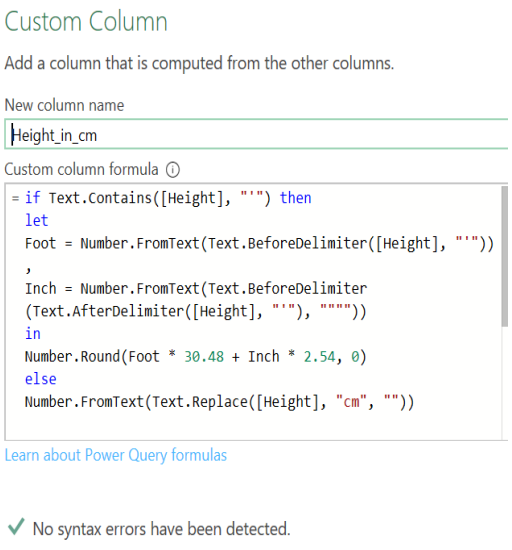
4. Normalizing Height and Weight Measurements

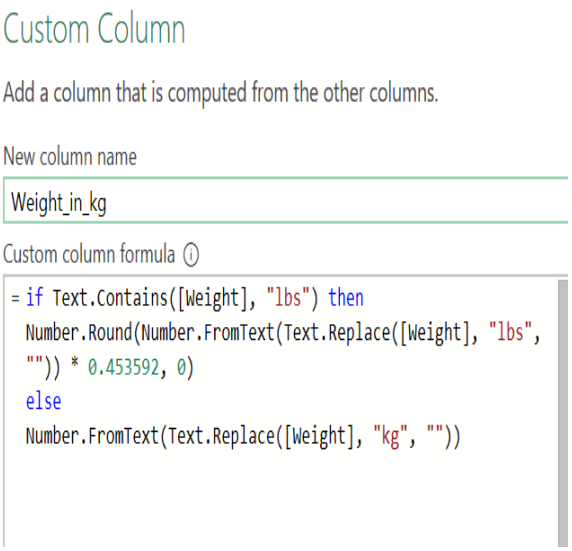
- **Issue**: Height and weight were recorded in mixed units (centimeters and feet for height; kilograms (kg) and pounds (lbs) for weight), complicating data consistency.

Below is a preview of the columns before cleaning:

Height	Weight
6'0"	185lbs
6'1"	179lbs
5'11"	170lbs
6'2"	196lbs
6'0"	172lbs
6'3"	203lbs
6'0"	183lbs
184cm	83kg
5'10"	168lbs
5'9"	161lbs
5'11"	146lbs
5'6"	130lbs
6'1"	190lbs
6'0"	172lbs
6'4"	174lbs
5'7"	148lbs
6'0"	165lbs
5'11"	172lbs
5'11"	161lbs
188cm	77kg

- **Solution**: Created a custom column with an IF statement to convert all height measurements to centimeters and all weight measurements to kilograms, ensuring uniformity across the dataset.





Height and Weight columns after Cleaning

BZ	CA
Height_in_cm	Weight_in_kg
170	72
187	83
188	87
181	70
175	68
184	80
175	71
191	91
178	73

Conclusion

The FIFA project showcased the importance of structured data cleaning and preparation in uncovering valuable insights. By leveraging Power Query functions in Excel, inconsistencies in the dataset were effectively resolved, resulting in a well-organized and reliable dataset. This

process highlighted how essential data preprocessing is to ensure the accuracy and relevance of analysis.

The cleaned dataset can now be used for various exploratory and analytical purposes, such as performance benchmarking, player scouting, and trend analysis within the FIFA dataset. This project also underscored the role of Excel's powerful tools in handling real-world data challenges, solidifying its relevance in the field of data analysis.

As a next step, this dataset could be integrated with advanced visualization tools like Power BI or used for predictive modeling to gain deeper insights into player performance.