First issue is having two datasets, so let’s check what’s up with them.

No difference in headers other than their organisation, so let’s go with v2, as if there are differences in any of the stats, can assume these are more up to date

Some notes on the abbreviations in the headers

*“W/F” stands for “Weak Foot”*

*“SM” stands for “Skill Moves”*

*“IR” stands for “International Reputation.”*

*As for "A/W" and "D/W" I would guess it stands for "Attacking Winger" and "Defensive Winger"*

**Choosing dependent variables**

Obviously, the two best dependent variables are Overall and Potential, as they describe ability.

Value and Wage could also be dependent, and there may be some interdependency between these four variables. I’d predict decent correlation between overall and potential, as well as value and wage. And again, between overall/potential and value/wage.

Release clause as well, just because it’s monetary

Some questions posed by the author:

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 been playing at a club for more than 10 years!*
4. **'Value', 'Wage' and "Release Clause' are string columns. Convert them to numbers. For eg, "M" in 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 highly valuable but still underpaid (on low wages)? (hint: scatter plot between wage and value)**

The ones in bold, I’d definitely like to do, italics I’m not that interested in.

My own question:

1. How much does position affect value and wage. I’d expect goal scorers to be paid higher than defenders.
2. For positions (will require dummy and grouping), think of some unique attributes that may affect their overall value. So, perhaps skill moves for forwards, pace for midfielders, height and weight for defenders?
3. How much does mentality affect wage? Seems maybe an interesting point…

Columns to drop, round 1! Fight!

ID

Long name

Photourl

Playerurl

Joined

Loan date end

Eyeballing

Some erroneous data in Goalkeeping. Most data is under 99, some are in the 400s, nothing in between. No, there are, they seem to go *almost* descending when they’ve got 3 digits, lowest is 231 maybe. Highest is 440.

Actually, narrowed the data down and spotted someone with a 112, as a striker. Seems strange, but most of the high vaues are GKs, so it might just be how they’re measured.

Round 2

Contract

Preferred Foot

Total Stats – A bit situational as some of the stats measure heavy in favour of certain positions

Base Stats – just don’t care

A/W – don’t know what this is

D/W – Attacking wing and defending wing? It’s so specialised I just don’t care.

**Variables to deal with**

Club – has a bunch of \n’s in front of the name

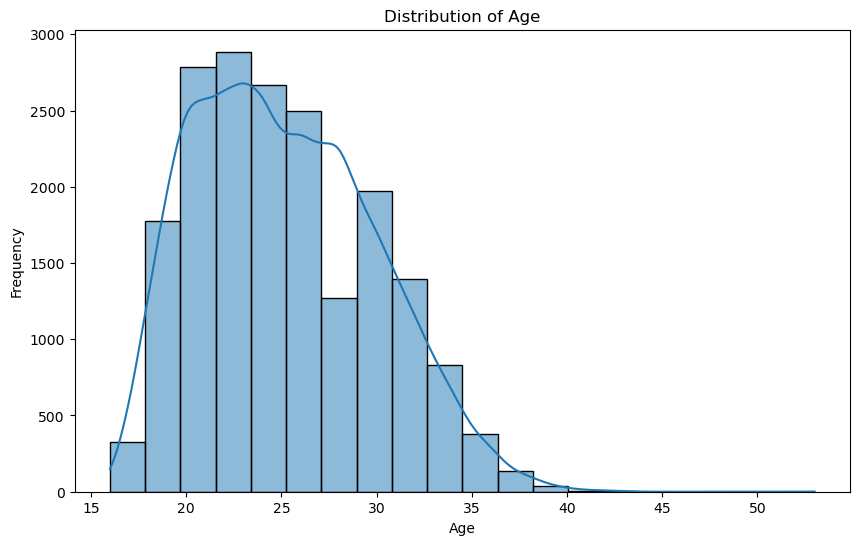
Positions – won’t drop this yet, but best position is certainly better for most obs… Could split these and create dummies for each position?

Height – ends in cm

Weight – ends in kg

Value – has EUR symbol, foats, and ends in M

**Baby’s First Python Histogram**



Creating a Histogram in Tableau

RC on the measure -> Create -> Bins