ADVANCED RULES – SUMMARY MATERIALS

GROUPING

By far the best reference for this is the pandas user guide, https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html

I would recommend reading that article in its entirety to understand the power of groups. For most of our work, you will likely want to focus on Transformation. The pattern you'll follow will be:

1. (Optional) Create a temporary helper column with the variable you want to group on or aggregate.
2. Use df.groupby('CHILD')['HELPER_COLUMN'].transform(<your function goes here>) to get a new column with the information you need.
    a. Some functions are transformers automatically, e.g. idxmin, idxmax, shift
3. Use this as extra information to build your error index.

MERGING

If you are already familiar with SQL and want to learn how to do the same thing in Pandas, there's a nice cookbook here https://pandas.pydata.org/pandas-docs/stable/getting_started/comparison/comparison_with_sql.html#compare-with-sql-join

Otherwise, a good reference is this section of the user guide https://pandas.pydata.org/pandas-docs/stable/user_guide/merging.html#database-style-dataframe-or-named-series-joining-merging

The steps you want to check when performing a merge are:

- Which dataframe is the primary dataframe (the one you will highlight the errors on)?
- Make sure you call `.reset_index()` on the primary dataframe to create an 'index' column.
- Do you need to know if there is no match on your primary dataframe (LEFT join) or is it fine to just check entries that are present in both (INNER join)
- Is the merge one-to-one or one-to-many? Think about how this may effect your output and the calculations you need to make?
- Then make the filtering calculations on the merged dataframe.
- To get the original index locations, use the index columns you made earlier, merged['index'].unique()

It's also worth considering your test layout. Phil's pull request from the video https://github.com/SocialFinanceDigitalLabs/quality-lac-data-beta-validator/pull/303 has a good test layout for more complex cases.

EXTERNAL DATA DOCUMENTATION

I have added external data documentation to the README of the repo – this is on the front page, under "The input data" https://github.com/SocialFinanceDigitalLabs/quality-lac-data-beta-validator

I'll try to keep this up to date if I add anything else – but this is a reference for where to find last years data, collection year info, etc.

EXERCISES

No exercises for this module – but I strongly encourage you to work through the videos slowly and follow along the code until you feel the concepts click.

When you do, pick up an advanced rule and have a go!