Population flow

Orders - original data
• Total: 3,421,083

orders_products_prior - original data
• Total: 32,434,489

Products - original data
• Total: 49,672

Customers - original data
• Total: 206,209

Orders - after consistency checks
• Total: 3,421,083

orders_products_prior - after consistency checks
• Total: 32,434,489

Products - after consistency checks
• Total: 49, 672

Customers - after consistency checks
• Total: 206,209

Orders_products_combined
• Total: 32,434,489

Orders_products_merged
• Total: 32,404,859

Orders_products_all
• Total: 32,404,859

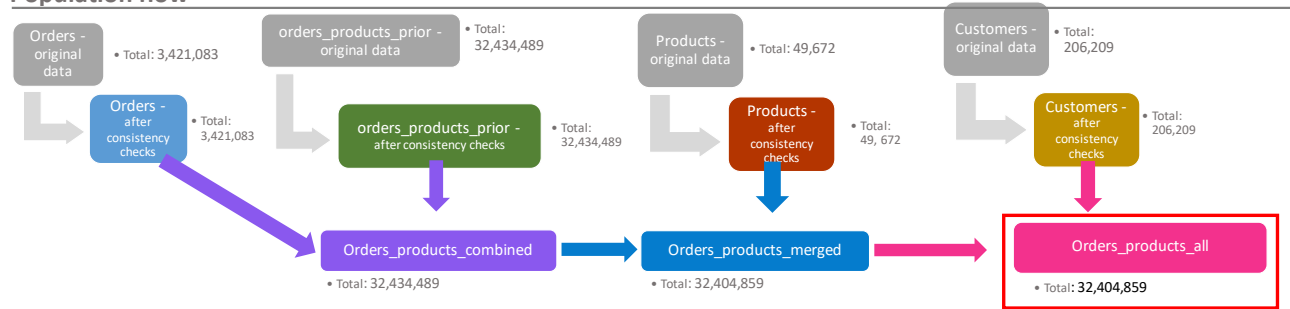**Exclusion flag**

Condition: max_order < 5
Obervations to be removed: 1,440,295
Final total count of order_products_all: 30,964,564

1.) The grey boxes in the first row of the population flow represent the original data sets as they were when first downloaded. The total fields add the count of the rows when imported into Jupyter.

2.) The second row of boxes (coloured) represents the data sets **after** manipulation, e.g., removed missing values and duplicates. The total fields sum the count of the rows after conducting these operations. This offers a visual oveview of how the data *flows* throughout the data consistency checks.

3.) The third row, where also the arrows are coloured, represents the merges performed between the datasets. The total fields sum the count of the rows in the merged datasets, ending with the final dataset (in the red box).