The Analysis of Commonly Used Transformations

Figure S1 shows the top 10 commonly appeared transformations in our dataset. Interestingly, creating tables by loading local files is the top transformation (i.e., *create_tables_load*). This is because *create_tables_load* appears in almost every wrangling script, which usually serves as the first step in data wrangling. *identical_operation* and *create_columns_mutate* are both ranked second. The *identical_operation* is often used as a prerequisite for the following transformations, such as *combine_rows_summarize*, *transform_tables_sort*, and *create_columns_mutate*, while the *create_columns_mutate* is often used to derive new columns by mutating existing columns, which is critical to preparing the desired data for subsequent analysis. As the raw data often contains some irrelevant or incomplete rows/columns, it is necessary to utilize the *delete_rows_filter* and *delete_columns_select_keep* transformations (ranked 4th and 5th, respectively) to filter them out. We believe that the analysis of commonly used transformations can guide the design of data transformation recommendation tools.

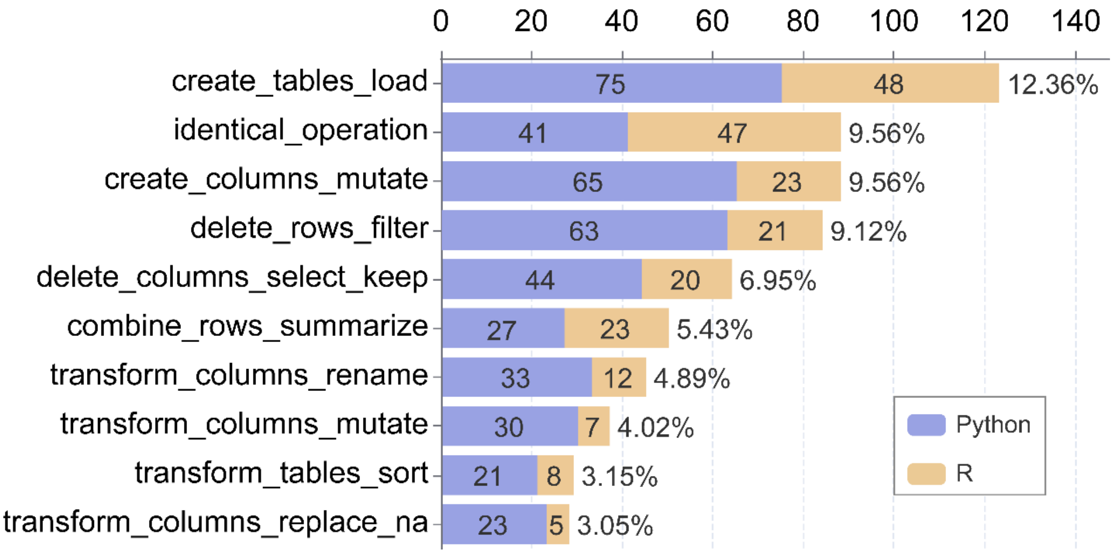

Fig. S1. Top 10 commonly appeared data transformations in our dataset. The unabridged frequency ranking of data transformation types is reported in the 2-transformation-type-results.csv.