Interesting! I tried to solve this problem a bit differently, as outlinedhere. The problem with the approach you've outlined is that pd.read_csv

will have to load the entire dataset into memory at full precision. Although you do get a reduction in memory usage after the transform in reduce_mem_usage

, the peak memory usage remains the same. Which means that you're still bottlenecked on memory (and glacially slow swap, if the machine has it enabled). My approach around that problem is to use converters to ensure that the data is converted to the more succinct dtype at load time. It isn't entirely free, the load time goes up by about a bit, but the memory usage never peaks.