**BACKGROUND:-**

I have a large amount of data in 5 separate spreadsheets – in total about 150,000 rows of data. Each row has up to about 21 columns of data.

Commonly, throughout the spreadsheet there is a huge number of cases where for example there might be, say, 4 to 12 rows of data in a spreadsheet where, say, there are two separate columns that can “jointly” define whether the rows are in fact duplicate in the sense that they relate to the same item. Equally there are many rows where there is only one row per item (versus multiple rows)

However, where these two columns can and do define the commonality of the rows relating to the same item, we can have a situation (using the example of say 12 rows thus relating to the one item) where there is a situation where there is the same or different data on these 12 rows appearing in different columns.

So just deleting duplicates is not an option per se because the rows, that would be automatically deleted using that process, contain data that would not appear in columns on the row that is retained during that process.

I am attaching a simplistic example of what I am talking about.

**THE OBJECTIVE:-**

The objective is to sort the spreadsheet initially (in this case by **Columns B and D**) in order to determine what rows are actually duplicates.

 That done, for each of these rows that relate to being a duplicate (of say the other eleven rows in my attached example), there is data on each row (in separate columns) that may be unique to the same item and it is a matter of **how do we capture that data on those “about to be deleted subsequent 11 rows” and put that on the first row of data to be retained?**

N.B. (now in my example spreadsheet attached I have *colour coded* similar *data just to highlight* the same data in different columns *for your benefit*). I can say this however – I simply really don’t know if the same data that is appearing in one column ***actually*** appears in other columns on different rows – in my spreadsheets (simply because there is too much to check).

Also for me I need to retain the data that appears in **Column P** (e-mail addresses) on the first row (that is the row we are retaining). In respect of all the other columns of data (excepting Columns **A, B, C, D and P** which **must** be retained) i.e., the data appearing in the other columns, the emphasis is on retaining ***as much as this as is possible*** so that ***as long as we have one item of such data in every column*** ***possible***(that is every column to which data is available) then that would be sufficient (even if we ended up with some duplicate data in different columns on the same row if we had to and even if we had situations where we ultimately lose some of that data in the delete duplicates process).

Obviously I need the answer to be such that it is **automated and** **time efficient** as I can’t manually go through 150,000 rows of data checking each and cut and pasting data from one row to another.

**So I need a step by step process to achieve these objectives** (with a clear explanation – aimed at a **non-technical** person - of what to do with each step).

I also need an indication (rough estimate) of how long you might think it could take to actually process this data according to the recommendations you make with respect to **the timely and efficient processing** of such duplicate data so we end up with an unique row of data for each item in my 5 spreadsheets of a total of approximately 150,000 rows of data in total. **The purpose of that question** is not only to validate **the timely and efficient** answer you have to this question but also to allow me plan for time required to do this.

In the event that there is not a timely and efficient process possible to solve this problem to my original satisfaction in terms of objectives, then I will need a solution that is the next best option to achieve the best possible solution in these circumstances.