Requirement

This is a mocked up interview exercise for manipulating a large dataset.

The client has accumulated test observations from various teams and operations. Each record is a single test observation. The observations are reported in ascending order by period, period 1 is the first observation, period 2 is the second, etc.

Collected columns:

* Report id – text
* Report period - number
* Test number – text
* Serial number of the tested device – text
* Type of the tested device – text
  + The client believes that there is only one distinct Type per test
* Size of the tested device – number
  + The client believes that there is only one distinct Size per test
* Start Length – number
* End Length – number
* Total Length Worked - number

The requirement is to aggregate this dataset and produce report following below rules:

Two or more observations should be aggregated if ALL below conditions are met:

* They have the same report id
* They are in sequential report periods
* Their test numbers are the same when all non-alphanumeric characters are removed
* The serial numbers are matching with each other when all non-alphanumeric characters are removed, and one serial number is a substring of another. For example, AX3072 will match to AX307 and AX3072B but not AXM301. A blank serial number matches all other serial numbers.
* All matches are case insensitive

The aggregated report will have below columns:

* Report id – text
* Test number – text
* Serial number of the tested device – text
* Type of the tested device – text
* Size of the tested device – number
* Min Report period - number
* Max Report period - number
* Start Length – number
* End Length – number
* Total Length Worked - number
  + The client believes this will reflect the difference between Max End Length and Min Start Length of aggregated observation