**Background**

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

Below is an imagined illustration of desired grouping for aggregation

|  |  |  |  |
| --- | --- | --- | --- |
| REPORT\_ID | REPORT\_PERIOD | TEST\_NO | SERIAL\_NO |
| 123 | 1 | 1A | AB |
| 123 | 2 | 1A | AB111 |
| 123 | 3 | 1A | AB111 |
| 123 | 6 | 2A | C22 |
| 123 | 7 | 2A | C22 |
| 123 | 8 | 2AD | E3 |
| 123 | 9 | 2A | F4 |

**Requirement**

In a database of your choice, create the data structure, build procedure to do the grouping and aggregation.

Document any issues found in the dataset, or false assumptions the client made.

Make recommendations for any adjustments to address these issues.

**Submit**

A csv of aggregated data.

All code including cleanup, manipulation and process of data.

A document outlining the difficulties encountered and any recommendations for changes.