CLEAN\_WQP\_DATA.R

**flag\_missing\_results** function:

**commenttext\_missing** = c('analysis lost', 'not analyzed', 'not recorded', 'not collected', 'no measurement taken')

wqp\_data\_out <- wqp\_data %>%

mutate(flag\_missing\_result =

( is.na(ResultMeasureValue) & is.na(DetectionQuantitationLimitMeasure.MeasureValue) ) |

grepl("not reported", ResultDetectionConditionText, ignore.case = TRUE) |

grepl(paste(**commenttext\_missing**, collapse = "|"), ResultCommentText, ignore.case = TRUE)

)

I do something similar in my workflow… I suggest hybridizing their approach with ours -essentially incorporating the columns they are using into our workflow (while not getting rid of our own), identifying the words to put into the grepl()’s by using MM’s word matrix function.

I added the flag\_missing\_results function to the no\_data\_samples step of the harmonize\_silica function. I filter out anything flagged by this.

MM matrix function to add?

**flag\_duplicates** function:

By default, a record will be considered duplicated if it shares the same organization, site id, date, time, characteristic name, and sample fraction. In my workflow, I don’t flag duplicates, I remove them, and my definition of duplicates is much more crude (for me, duplicates are defined as having identical values in all columns that remain after our initial renaming/selecting step at the very beginning of the workflow). I suggest using their approach instead of mine.

FORMAT\_COLUMNS.R

**ResultDetectionConditionText,**

**DetectionQuantitationLimitMeasure.MeasureValue,**

**ResultDetectionConditionText,**

**ResultCommentText**

These columns… do we get rid of them? If so, we shouldn’t! In fact, moving forward I suggest keeping all columns from the WQP pull.

**drop\_vars:** Same comment as above… We do in fact drop variables and we also rename the ones we keep. Their default is to keep all columns. We might not want to drop any columns like we have been doing.

**vars\_to\_numeric:** We do something similar where we create a new column that identifies ResultMeasureValue (for us we rename this to value) fields that are numeric. They also do this for DetectionQuantitationLimitMeasure.MeasureValue, which we should also do!