Data Wrangling Report

The cleaned data can be found at:

https://drive.google.com/drive/folders/1CD8-oHrF6VQyjEdSL8v6PSWbstwB0T08

Steps taken:

1. Combined data from 2013 – 2017 into one data frame.
2. Delete unnecessary columns
   1. “X.id” – only URLs
   2. “sample.samplingPoint” – only URLs
   3. “codedResultInterpretation.interpretation” – NULL column values
   4. “sample.samplingPoint.notation” and “sample.samplingPoint.label” - the names of the locations have too many levels to make any meaningful analysis.
   5. "determinand.notation" - removed because it's redundant information: it's a numeric code assigned to each type of determinand, which is already given unique names in "determinand.label".
3. Simplify column names

|  |  |
| --- | --- |
| **Original Column Name** | **New Column Name** |
| sample.sampleDateTime | time |
| determinand.label | \*unchanged\* |
| determinand.definition | \*unchanged\* |
| resultQualifier.notation | resultqualifier |
| result | \*unchanged\* |
| determinand.unit.label | resultunit |
| sample.sampledMaterialType.label | materialtype |
| sample.isComplianceSample | compliance |
| sample.purpose.label | purpose |
| sample.samplingPoint.easting | easting |
| sample.samplingPoint.northing | northing |

1. Convert “time” values to POSIXct format
2. Convert columns to factors as needed
3. Filter dataset for only the most frequently occurring tested determinands.