**Summary of Flagged Data and Potential Fixes***\*logbooks = samples,  
\*length frequencies = length frequencies  
\*biological samples (details) = Fish Details*

**Fish Details: gaspereau\_fish\_details.csv**

|  |  |  |  |
| --- | --- | --- | --- |
| **Flag Name** | **Number of Occurrences** | **Description** | **Potential Fix** |
| FLAG\_NO\_MATCHING\_SAMPLE | 6876 | No sample with the same date and site number as the fish details | Please import with ghost sample and have some system in place for replacing a ghost sample with the proper sample.  I think samples (logbooks) can be found for some of these occurrences, but it will take time to locate the samples between Science and Stats. |
| FLAG\_SITE | 1922 | Sites with names instead of ids, multiple sites, no site id, or alternate nomenclature | **CONVERT E.MACFARLANE, Eric Mac, Eric MacFarlane, J McFarlane to John Eric MacFarlane.**  **JA Coady keep as is**  **Multiple sites, keep as is.   I see 118 samples from 1989 that are associated with ‘blank’. Can you convert ‘blank’ to Unknown and upload.** |
| FLAG\_AM\_PM\_PERIOD | 118 | AD instead of AM or PM | **AD = All Day**  Valid category, should remain as is. **Period will therefore have 3 options; AM, PM, AD** |
| FLAG\_SEX | 2 | ‘B’, ‘A?’ | convert ‘B’ and ‘A?’ to null |
| FLAG\_MATURITY | 3 | 0, 44 | **44 = 4**  **0 = null** |
| FLAG\_FSP\_1 | 1 | 33 | **convert 33 = 3** |
| FLAG\_FL\_STD | 264 | off by factor of 10 (cm vs mm conversion?) | multiply by 10; YES  **lengths were entered as cm, but should be mm** |
| FLAG\_FL\_WET\_FROZEN | 5 | ~~off by 10x or 100x~~ | **FL\_Frozen and FL\_STD were both updated with the correct values in the csv.** |
| FLAG\_WEIGHT\_OUTLIER | 6 | weight greater than 500g | **Correct weight is 193.4g for the major outlier. Keep the other 5 weights above 500 as is. Those weights are associated with longer fish, so the weight makes sense.** |
| FLAG\_GONAD\_OUTLIER | 7 | gonad weight greater than 200g (all have gonad weight greater than total weight) | convert to null; YES |
| FLAG\_LEN\_WT\_RATIO\_OUTLIER | 33 | 4 standard deviations outside predicted weight based on length (linear regression) | leave as is |
| FLAG\_MULTIPLE\_SAMPLE\_POSSIBILITIES | 1034 | multiple samples with the same date and site location as the fish details | **FD - 2014/05/30 - #47 or 62 = #62**  I can’t fix the rest until I find their hardcopies. Please import with ghost sample. |
| FLAG\_AM\_PM\_PERIOD\_DISCREPANCIES | 196 | different AM/PM period versus matching Length Frequencies data | **LF- 2018/06/23 - #41 = PM**  **FD - 2014/06/10 - #47 = PM**  I don’t have the hardcopies for the remaining discrepancies to check. Please NULL the AM/PM field for those. |
| total number of rows in the table | 36912 |  |  |

**Length Frequencies: gaspereau\_length\_frequencies.csv**

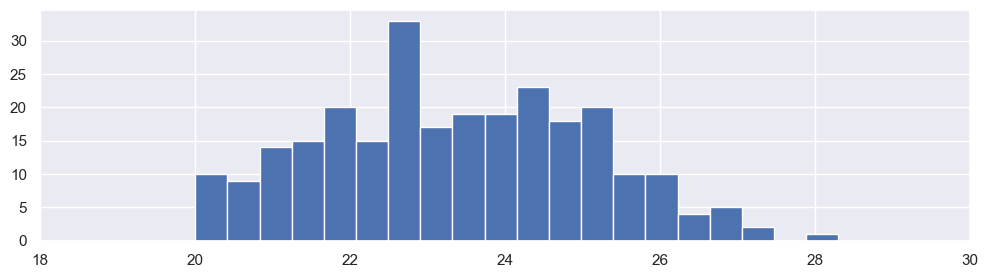
|  |  |  |  |
| --- | --- | --- | --- |
| **Flag Name** | **Number of Occurrences** | **Description** | **Potential Fix** |
| FLAG\_NO\_MATCHING\_SAMPLE | 1954 | No sample with the same date and site number | **LF – 2014/05/21 - #21 = #41**  Please import the rest with ghost samples |
| FLAG\_SITE | 627 | Sites with names instead of ids, multiple sites, no site id, or alternate nomenclature | **Import site names as they are. Will have to be fixed later.** |
| FLAG\_AM\_PM\_PERIOD | 47 | AD instead of AM or PM | **AD is legitimate. Stands for All Day. Keep as is.** |
| FLAG\_AM\_PM\_PERIOD\_DISCREPANCIES | 69 | different AM/PM period versus matching Fish Details data | **(repeat) FD - 2014/06/10 - #47 = PM**  **(repeat) LF – 2018/06/23 - #41 = PM**  I don’t have the hardcopies for the remaining discrepancies to check. Please NULL the AM/PM field for those. |
| total number of rows in the table | 11440 |  |  |

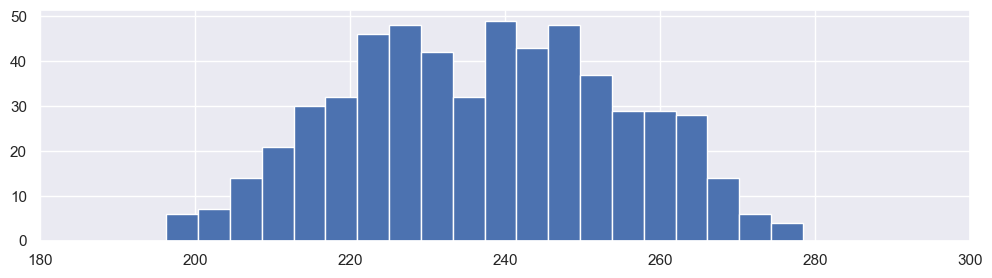
**Sample Data (logbook): gaspereau\_sample\_data.csv**

|  |  |  |  |
| --- | --- | --- | --- |
| **Flag Name** | **Number of Occurrences** | **Description** | **Potential Fix** |
| FLAG\_DATETIME | 31 | entry has missing date | Site 48, 2019 – did not fish  Site 51, 2017 – did not fish  Okay, for the samples flagged here, I suspect the ones entered from 2006 to present all represent years where that trap was not fished at all during the season. These can all be removed.  The information from site 37 in 1988 should be kept as it does appear that the trap was fishing that year. |
| FLAG\_HOURS\_FISHED | 28 | hours\_fished are noted as ‘maximum ’ | Is it possible to import this data with MAX listed under hours fished? I am waiting for answers from others to try to determine what MAX can mean based on the commercial fishing rules. Eventually I would like to go in and switch MAX to a numeric value, but I just need more time to determine the correct numeric value. |
| FLAG\_SITE | 23 | site labelled '1A,8' | Please leave as is for now. Will try to correct when hard copy is found. |
| FLAG\_AM\_PM\_PERIOD\_DISCREPANCIES | 4 | different AM/PM period in Fish Details versus matching Length Frequencies data | I could not find the 4 instances showing TRUE. I’m not sure if maybe I corrected them already? |
| total number of rows in the table | 15213 |  |  |

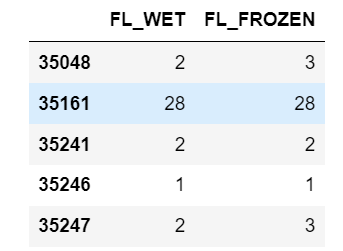
**Descriptions of Flagged Data Issues Discussed in the Meeting**

* **Fish Details**
  + No matching Sample: **FLAG\_NO\_MATCHING\_SAMPLE**
    - No sample with the same date and site number
  + Sites with names instead of ids, multiple sites, no site id, or alternate nomenclature: **FLAG\_SITE**
  + PERIOD: **FLAG\_AM\_PM\_PERIOD**
    - AD -> AM?
  + AM/PM discrepancies vs same sample in Length Frequencies: **FLAG\_AM\_PM\_PERIOD\_DISCREPANCIES**
  + SEX: **FLAG\_SEX**
    - B, A?
  + MATURITY: **FLAG\_MATURITY**
    - 0, 44
  + FSP: **FLAG\_FSP\_1**
    - 33
  + Length
    - FL\_STD: **FLAG\_FL\_STD**
      * off by factor of 10 (cm vs mm conversion?), dates: 2019-05-31 - 2019-06-21





* + - FL\_WET and FL\_FROZEN: **FLAG\_FL\_WET\_FROZEN**
      * 5 entries, off by 10x or 100x



* + - Imputing values for FL\_STD using FL\_WET or FL\_FROZEN
      * 1457 entries
      * Factor of approximately 1.033 (CONFIRM)
      * This conversion will occur in dm\_apps based on whether the inputted fish is fresh or frozen
  + Outliers
    - Weight: **FLAG\_WEIGHT\_OUTLIER**
    - Gonad weight: **FLAG\_GONAD\_OUTLIER**
    - Length to weight ratio: **FLAG\_LEN\_WT\_RATIO\_OUTLIER**
      * 4 standard deviations flagged,
      * Note: 3 stdev would result in 101 flagged entries
  + Ambiguous Sample:  **FLAG\_MULTIPLE\_SAMPLE\_POSSIBILITIES**
    - Multiple sets of fish details at the same site on the same date
* **Length Frequencies**
  + No matching Sample: **FLAG\_NO\_MATCHING\_SAMPLE**
  + Sites with names instead of ids, multiple sites, no site id, or alternate nomenclature: **FLAG\_SITE**
  + River and location (MARGAREE / LOWER) will be added to Sites table
  + Period: **FLAG\_AM\_PM\_PERIOD**
    - AD -> AM?
  + AM/PM discrepancies vs same sample in Fish Details: **FLAG\_AM\_PM\_PERIOD\_DISCREPANCIES**
* **Samples**
  + Entries have missing dates: **FLAG\_DATETIME**
  + Entries for hours\_fished are ‘maximum ’: **FLAG\_HOURS\_FISHED**
    - Can this be converted to a numeric value?
  + Sites with names instead of ids, multiple sites, no site id, or alternate nomenclature: **FLAG\_SITE**
  + Some LF and FD are not matched to any samples
    - Create “ghost samples” for any samples after this step that remain unmatched
      * All length frequencies and fish details should be imported to dm\_apps
    - This step will be left until clarifications from this data request are complete to maximize matches
* **Samplers**
  + “TrapSupervisors”
  + Some duplicate names, whitespace, alternate spellings
  + May be linked to Sites
* **Sites**
  + No lat long data or descriptions
    - Will leave blank in dm\_apps for now; can update as information becomes available
  + Site should include location (Lower / Upper) from Length Frequencies (Similar to ‘zone’ from Samples)
    - Sites 12 and 62 have ambiguous zones
      * In Samples:
        + Site 12

Lower (1180 occurrences)

Upper (29 occurrences)

* + - * In Length Frequencies
        + Site 12

Lower (622 occurrences)

Upper (12 occurrences)

* + - * + Site 62

Lower (12 occurrences)

Upper (278 occurrences)

* + - * Assumptions to **CONFIRM**:
        + Site 12 = Lower
        + Site 62 = Upper
  + Some names are included instead of site number
    - None of these unverified sites exist in Sample entries, and therefore are not currently imported into dm\_apps… at least before introducing ghost samples (described above).
* **Gear description for Gear 81**
  + ID = 81
  + Description =
  + Gear Code =
  + ISSCFG Code =
  + NAFO Code =
* **Fish Details - Maturity methodology changes over the years**

