Purpose: working framework to the DS health and conditions dataset EDA

1. Begin EDA
   1. Explore dataset
      1. Metadata is sparse: fill out the metadata tab of the dataset for more context
         1. How many NAs per variable?
         2. What can be potential response vs predictor variables?
            1. What does the response variable represent?

Greater fish health equates to what? Greater growth? Greater survival?

* + - * 1. What can be used as a predictor variable?

Are there predictors that represent “good conditions” that “benefits” DS?

* + - 1. Correlation matrix between to response and to other predictors
    1. Maximizing usable data:
       1. Test some thresholds, only accept variables with enough data?
       2. Maximize correlation and minimize Nas
    2. What does the response variable represent?
       1. Greater fish health equates to what? Greater growth? Greater survival?
       2. Which variable is the best candidate?
          1. Minimize NAs, outliers
          2. Normalizable distribution
    3. What can be used as a predictor variable?
       1. Are there predictors that represent “good conditions” that “benefits” DS?
       2. Can I supplement predictors with other data, i.e., data from other surveys or continuous sensor data, to fill in NAs?
       3. Relationship to the response: plot it
  1. Preliminary models
     1. Highly dependent on what the response looks like
     2. Start with a correlation analysis
        1. 0.65 threshold for collinearity