The given raw dataset is in the CSV format with 1.16 GB storage. There are 3123117 observations with 58 variables, including the numerical response CPHL. We aim to use the remaining dependent numerical variables and their corresponding classifications to predict the value of CPHL, which is the measurement of phytoplankton concentration.

It can tell from the dataset that there are seven distinct gliders recorded data from seven different periods between 2013 and 2015, while the gliders were placed at various depths, latitudes, and longitudes. Nevertheless, the shortage of raw data is noticeable. A large number of data at around 19% overall is missing that spreads unevenly across the variables. Take one variable, VCUR, which is the value of seawater velocity at northward for example, a significant amount at 97% of the data is missing. It is crucial to balance the importance of missing data and the information it has contained during data preprocessing. Another difficulty of the dataset is the interaction between different variables. The giant size of variables will quickly impact the accuracy of correlation detection, so that it is vital to exclude the noise from disturbing features in data analysis.

We intend to apply Python and R on the project, while Python will mainly concentrate on missing data filling, and R will focus more on modeling and diagnostic. Our first step is to visualize the data and clear out the invalid value range. It is helpful to plot the correlation matrix with the library("corrplot"), scatterplot with a smooth spline, histogram by R to evaluate the potential relationship between predictors and response, and the percentage of the data quality (i.e., good or bad data). Different methods will be tested via python to deal with missing value, including mean, mode, KNN, random forest filling ways with varieties of sklearn libraries. After comparing the replacement accuracy, the best performance approach can be taken to acquire the completable without missing value. Then it goes to the other essential process that is the data modeling and algorithms. Due to the positively distributed numerical response CPHL, it is highly like that the topic is a regression model issue. In that case, we would like to test different models' performance, including xgboost, random forest, elastic net with python with higher efficiency, but mainly concentration the work on regression ones contains logistic, lasso, and support vector regression in R due to the precision. Besides, we also have the hyperparameter tunning plans to improve the accuracy after the comparison of measurements such as MSE, R-squared, and AUC-ROC diagrams and Anova tables.

The above processes on the modeling build will be assessed in R and Python to take advantage of both, and a better result with higher performance one can be selected.