WEEK 6

Merging datasets:  
Analysed code used to merge the ARTwarp data (Hawaii data Emily has been working on) and the ROCCA data (encounters pulled from OCEAN). The original code matched the data up by Sourcename, but I changed the code to match up by encounter ID instead.

Next, we need to fit a model to this merged data.

Remove variables with multicollinearity:

1. Remove highly correlated variables

* Calculated the correlation matrix to measure relationships between variables.
* Filtered for absolute correlations > 0.9 to detect high multicollinearity.
* Matched the row-column index values to actual variable names.
* Selected one variable from each pair to remove, prioritizing interpretability.

A heat map with red and blue squares

AI-generated content may be incorrect.

A table of numbers with black text

AI-generated content may be incorrect.

|  |  |  |
| --- | --- | --- |
| Highly Correlated Variables | Variable to Remove | Reason |
| FREQQUARTER3 & FREQMAX | FREQQUARTER3 | FREQMAX captures max frequency, more standard |
| FREQSTDDEV & FREQMIN | FREQSTDDEV | FREQMIN is a direct measure, easier to interpret |
| FREQRELBW & FREQMIN | FREQRELBW | Redundant with FREQMIN |
| FREQMAXMINRATIO & FREQMIN | FREQMAXMINRATIO | FREQMIN is more direct |
| DCQUARTER3MEAN & FREQCENTER | DCQUARTER3MEAN | Less commonly used metric |
| FREQMEDIAN & FREQCENTER | FREQMEDIAN | FREQCENTER is a better central tendency measure |
| FREQQUARTER1 & FREQCENTER | FREQQUARTER1 | Overlaps with FREQCENTER |
| FREQQUARTER2 & FREQCENTER | FREQQUARTER2 | FREQCENTER is a better summary statistic |
| FREQQUARTER3 & FREQCENTER | FREQQUARTER3 | “ |
| FREQMAXMINRATIO & FREQRELBW | FREQMAXMINRATIO | Already removing FREQRELBW |
| FREQQUARTER3 & FREQQUARTER2 | both |  |
| FREQSLOPERATIO & DURATION | FREQSLOPERATIO | DURATION is easier to interpret |
| NUMSWEEPSFLATUP & NUMSWEEPSFLATDWN | NUMSWEEPSFLATUP | NUMSWEEPSFLATDWN is likely more relevant |
| INFLMEANDELTA & INFLSTDDEVDELTA | INFLMEANDELTA | INFLSTDDEVDELTA captures variability better |

1. Eliminating Constant and Low-Variance Variables

* Columns that contain only one unique value across all observations are removed, as they provide no useful information for prediction.

Assessing Feature Importance

* A generalized linear model (GLM) is fitted to the dataset.
* The importance of each explanatory variable is calculated, and variables with very low importance (below a chosen threshold) are removed.

1. Manual Removal of Irrelevant Variables

Specific variables (Encounter, EncounterID, and Recording) are manually removed because they do not provide predictive value for the model.

1. Calculating VIF values for variables

When trying to calculate the VIF values, I get the error that contrasts can only be applied to factors with 2 or more levels.

Once this issue is fixed and we decide on the relevant variables to be predictors in the model, we can fit the model.