**Data & code changes log**

| **Date / time** | **Group member** | **File(s) added or edited** | **Notes** |
| --- | --- | --- | --- |
| 4/4 11:30pm | Anna | cleaning\_var\_selection\_v1.r | * The data is a time series, with 30 participants and 4 visits each (27 people only had 3 visits)   + Can we ignore this? Perhaps consider only using data from one of the visits for each participant? Not sure. * Created binary indicator variable for depression, 7 participants showed signs of depression for at least one of their visits * VIFstep on numeric variables showed multicollinearity problem with 3 predictors (using stricter threshold) * 6 predictors had missingness; doesn’t seem to be a huge issue for this dataset * Suggestion: be more selective about which predictors we consider, maybe choose 5-10 predictors max (could combine some metrics) |
| 4/9 8:20pm | Anna | cleaning\_var\_selection\_v2.r  clean\_vars.csv | * clean\_vars.csv was created using cleaning\_var\_selection\_v2.r. It contains a selection of 45 predictor variables and the outcome variable. 13 variables were removed as described in the [Predictors](https://docs.google.com/document/d/1Imy3F2QSr2rMD2Ri8P9_GqUj1hzBycRCgkaZ2HeloWE/edit?usp=sharing) doc.   + No missingness: columns with lots of missingness were excluded, and any other missing values were imputed from other visits of the same participant   + Duplicated the 3rd visit for 3 participants who did not have a 4th visit   + Eliminated multicollinearity among quantitative variables * Ready to make initial models using clean\_vars.csv |
| 4/12 | Caroline | binary\_log\_regression\_v1.R | * Warning messages: fitting binary logistic regression model, diagnostic plots? * Removed column with original depression total score * Fit two binary logistic regression models: cleaned data, cleaned data without outliers   + Created train (70% of observations) and test sets * Are we more concerned with sensitivity or specificity?   + Is it more important to predict if people have signs of depression (sensitivity) or if they do not have signs of depression (specificity)? * Model 1: AUC is ~0.7, so the binary classifier performs better than random guessing (0.5), but we may be able to find a better model |
| 4/10 | Hannah | data\_exploration\_trees.R | * Created dataset with depression\_total\_score removed because of correlation to depress\_binary → remember in future analyses to remove depression\_total\_score * Warning message (when creating tree object):   + In tree(as.factor(depress\_binary) ~ ., data = train) :   NAs introduced by coercion   * Best tree size as determined by running cvtree 50 times is 2 — variable that splits into terminal nodes is “anxiety\_perception” * When viewing the confusion matrix for the model fit on the test dataset, we notice it does not contain anyone who actually has depression (ie no true positive). Because our dataset overall is small, this is a motivation to not use train / test and move towards k-fold cross validation. |
| 4/12 | Caroline | svm\_v1.R | * Tried to fit SVM for diff kernel sizes, used the same train and test sets as the binary log regression * There is an error when trying to predict values using the train set |
| 4/24 | Caroline | cleaning\_var\_selection\_v3.R  clean\_third\_visit | * Did this together * Cleaned, chose 3rd trial * Did not do variable selection yet (bc that’s for lasso regression) |
| 4/26 | Caroline | cleaning\_third\_visit\_v2.R  clean\_third\_visit\_v2.csv | * Took out predictors that should not be used for regression * Did cleaning for 999 values and NA values |