This project started with a raw data set from the UCI Machine Learning Repository and a simple objective that calls for a systematic evaluation to relate measured and categorical data to one particular numerical data field.

Exploratory data analysis was presented in a separate notebook and began with an examination of the categorical data field for specimen gender. Every observation was reported as having one of three genders: male, female or infant. No material changes were needed here although there was a recognition that 'infant' samples are not necessarily young samples but rather samples lacking or having a neutral a sex. EDA was also done to trim the data set based on sanity checks. These checks looked for strange outlier data based on simple relationships between measurements in millimeters or grams. The data description furnished by UCI provided some guidance for this process along with basic expectations such as the requirement that a whole sample can not be heavier than its components. Additionally, the ratios between measured values and ring counts at every life cycle phase were deemed plausible.

Later, the subject of gender was revisited to look at Pearson correlation coefficients between the ring count and various measurements for each of the three reported genders. Randomization was done 10,000 times over to determine the p value for the observed dissimilarity between the sexes. The conclusion is that sex in abalone shellfish is more than a social construct. The fastest growing, most robust cases are female while the samples with no sex at all take on age with the least growth. Males are in the middle. The overall result was to retain sex for ring count projections.

With the data set deemed ready for implementation, one hot encoding was used to represent male and female cases with a '1' and '0' and neutral observations with a pair of 0's. (A pair of 1's is the only impossible combination here because an abalone can not be both male and female.) Regressions and machine learning classifications were used as a final sanity confirmation to show how each of the data fields that are input into ring count calculations relates to the other data fields used in this process. As with all other regressions and classifications, random forests were used with Bayesian parameter tuning and a 2/3 to 1/3 training to testing split.

A classification of sample maturity (as defined above) was done based on all the other data fields as modified. This classification was found to be prone to false positives to be reduced by padding with additional rings in the projection. A preliminary answer was obtained by buffering for a one-sided 95% confidence interval as the most pessimistic likelihood for any given specimen. From a classification matrix, the trade-off due to the large number of samples having to be culled advised the removal of one ring from the buffer. The ROC curve and its associated calculations along with the confusion matrix further shed light on these trade-offs and the limitations on the insight to be gained from the available data. Buffering was examined with the help of regressions.

A follow-on project would work these comparisons of calculated and actual ring counts into an index for quantifying the overall health of samples based on how rapidly they grow.