Isaac Connelly & Caijun Qin

December 6, 2021

Dr. Athienitis

STA 4504 Missing Data Project

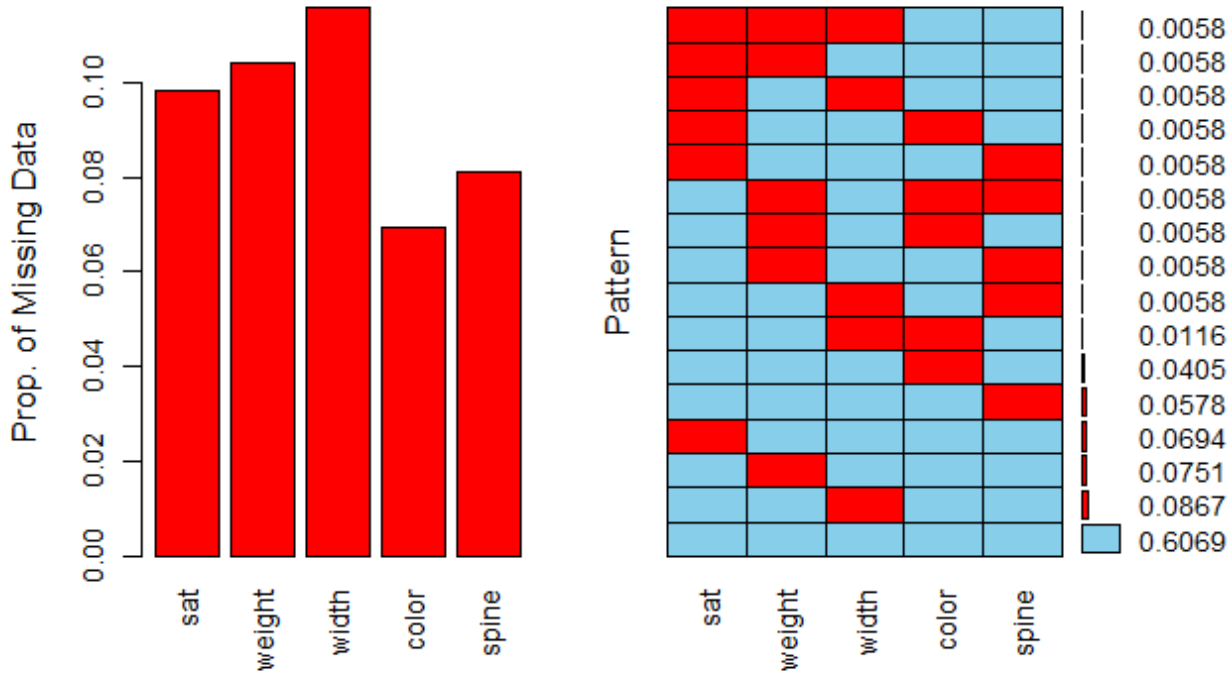
1. Introduction

For this project, we are working with the crab dataset which originally holds the number of male satellites (*sat*) as the response variable given *weight*, *width*, *color*, and *spine* of each female crab in the study. *sat, weight*, and *width* are nonnegative numerical variables, and *color* and *spine* are categorical variables, albeit being ordinal in nature.

Missing values exist throughout the data and will affect the quality of statistical analyses on this particular dataset. Based on descriptive statistics of missing data, *width* holds the largest proportion of missing values (> 10%) over all variables (Figure 1). In contrast, *color* holds the lowest proportion of missing data at ~ 7% (Figure 1). Proportion by pattern of missing data constitutes another perspective of descriptive statistics for the dataset. In Figure 2, a pattern constitutes a specific combination of five (5) Bernoulli trials corresponding to variables, with each trial being either a success (observed value) or failure (missing value) for the specific variable. For example, the proportion of data missing exactly *sat*, *weight*, and *width* is 0.0058 (top row), and the proportion of data without missing values for any variable is 0.6069 (bottom row) in Figure 2.

The specific counts of missing values per variable are given in the order for c("sat", "weight", "width", "color", "spine"), which are c(17, 18, 20, 12, 14). Critically, all variables have a missing proportion of > 0.05; a general rule of thumb for data cleaning proposes the removal of any variable with > 0.05 of its values being absent. Removing all variables would cause drastic detriment to statistical analyses, as having some observed data with missing values is more viable than no data. From Figure 2, the pattern of

missing data does not appear to hold consistency across variables. The proportion of observations missing exactly two (2) variables is approximately 10-fold or more compared to the proportion of observations missing three (3) or more variables. From Figure 3, the proportion of missing data conditional upon the count of satellites appears to fluctuate rather than remain consistent. Therefore, we confidently reject the notion of *missing completely at random (MCAR)*. We cannot truly prove whether the data is *missing at random (MAR)* or *missing not at random (MNAR)* due to missing values being inherently unobservable for testing, but the assignment implies MCAR.



Figures 1-2. Left: Proportion of observations with missing values (NA) per variable. Right: Proportion of different patterns of missing data.

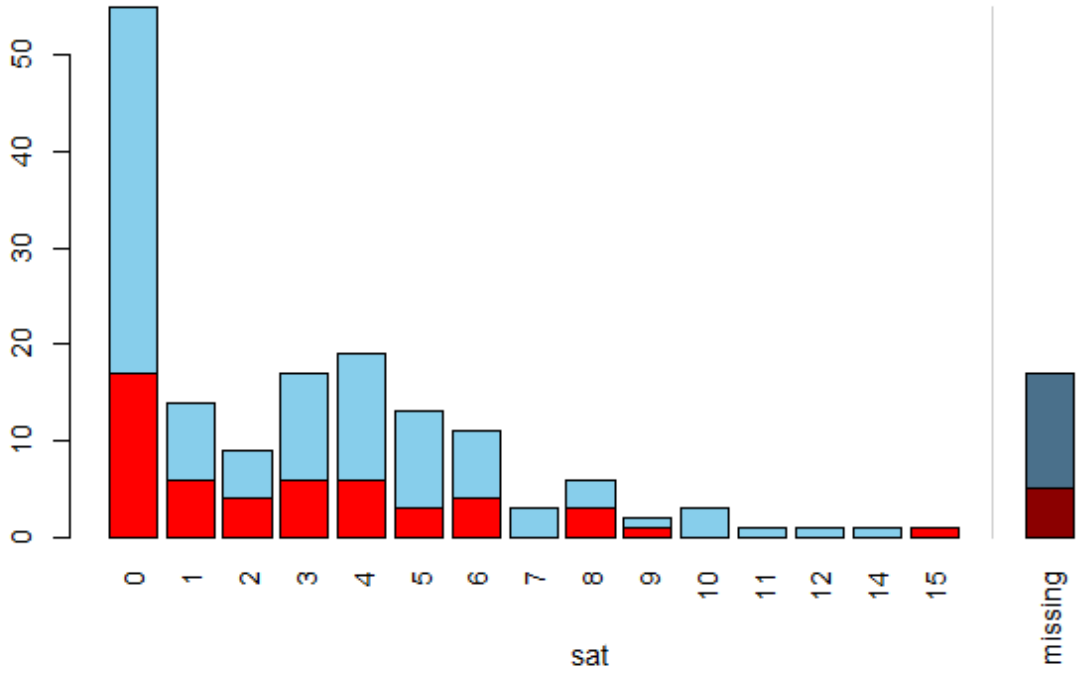
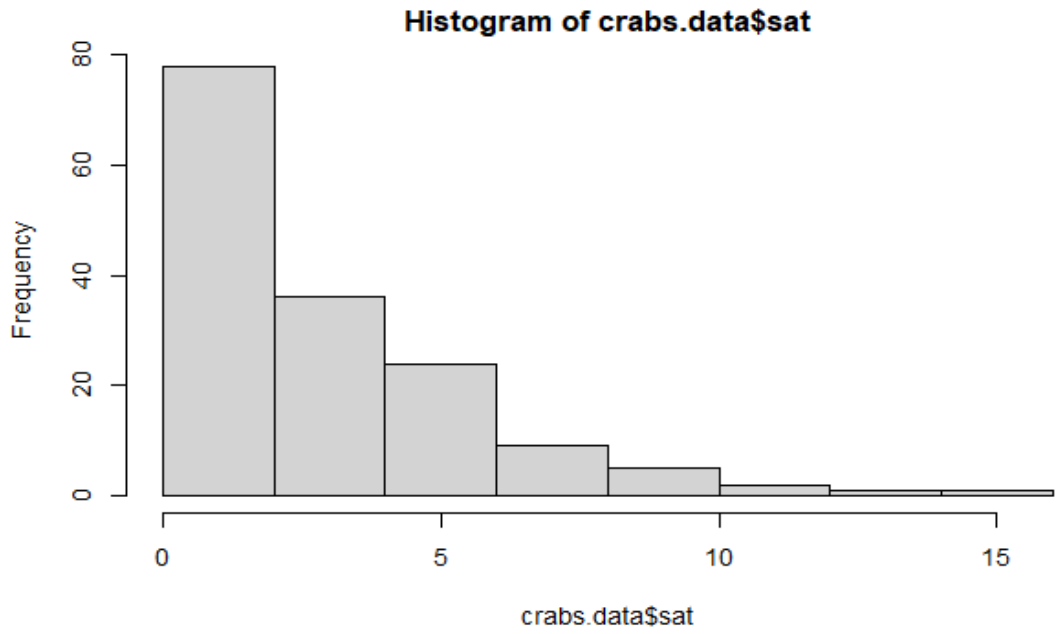
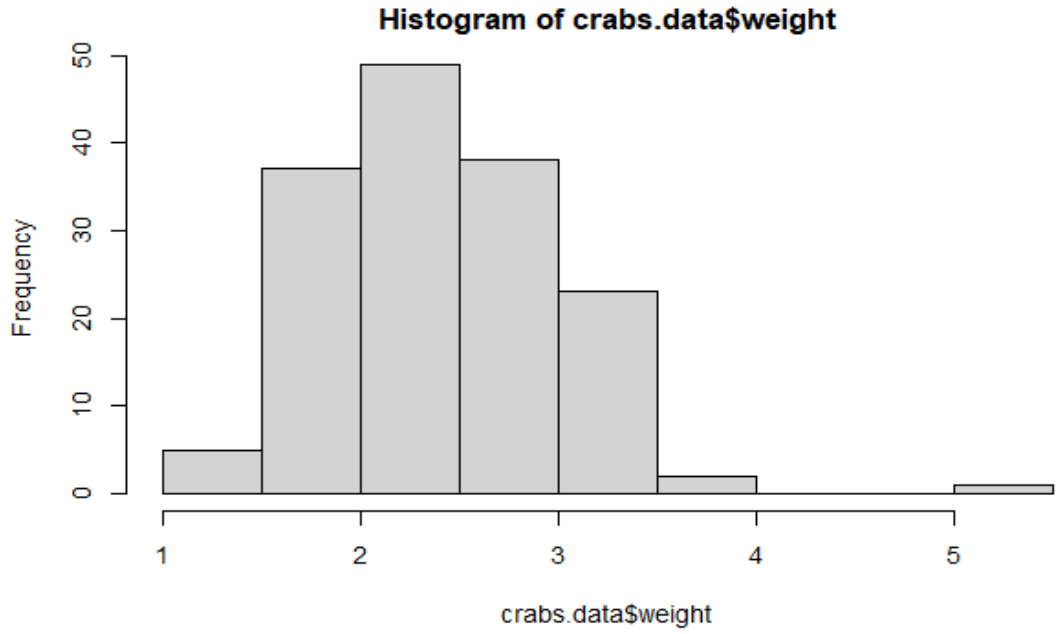


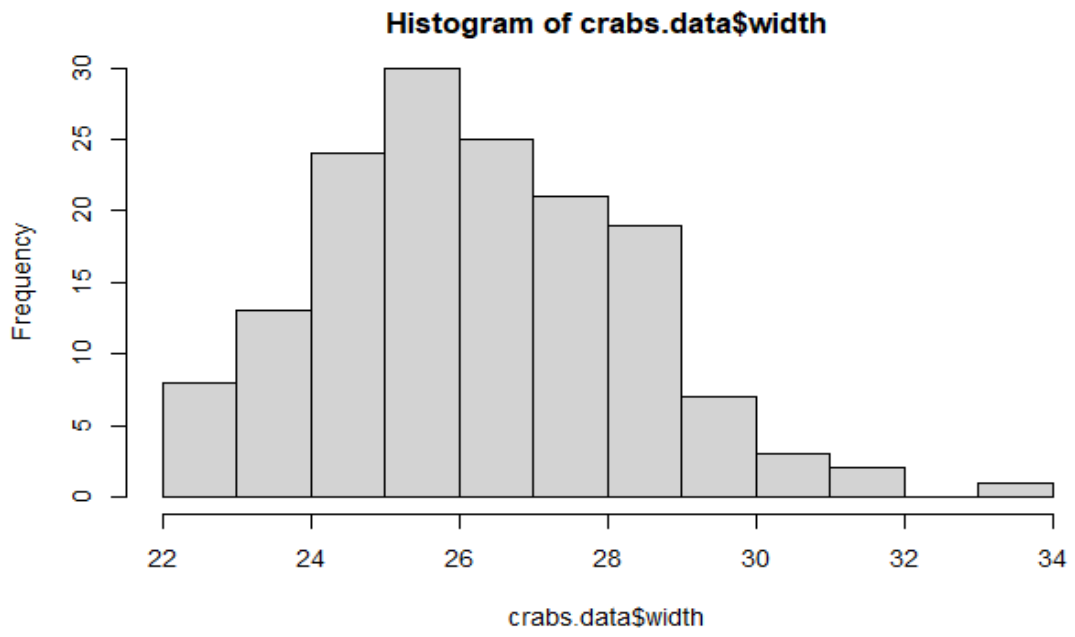
Figure 3. Proportion of missing data based on number of satellites.

2. Choosing a suitable imputation method for each variable proved to be a cumbersome task. We conducted a brief study on the observed data with descriptive statistics. For each numerical variable, a histogram plots the binned density of observed values under the variable. For each categorical variable, a pie chart plots the category proportions of observed values. From initial glance, observed values for *weight* and *width* seem to be normally distributed (Figures 5 and 6). Hence, the imputation method used for them is "norm". *sat* seems to follow an exponential distribution (Figure 4), but we do not know what the true population distribution looks like. Having also tried "norm" on *sat*, negative values appeared as imputed values. We plan to use Poisson with a log link function to model satellite count data using a generalized linear model (GLM), which does not accept negative values in the response variable. We then finalized on "sample" as the imputation method for *sat*. Since all observed values from *sat* are nonnegative, this method would never impute a negative value. Furthermore, using sampling is appropriate given that the exact distribution is uncertain. For both categorical variables, the "rf" (random forest) method is used for the suitability for classification. Although "cart" could also be a worthy

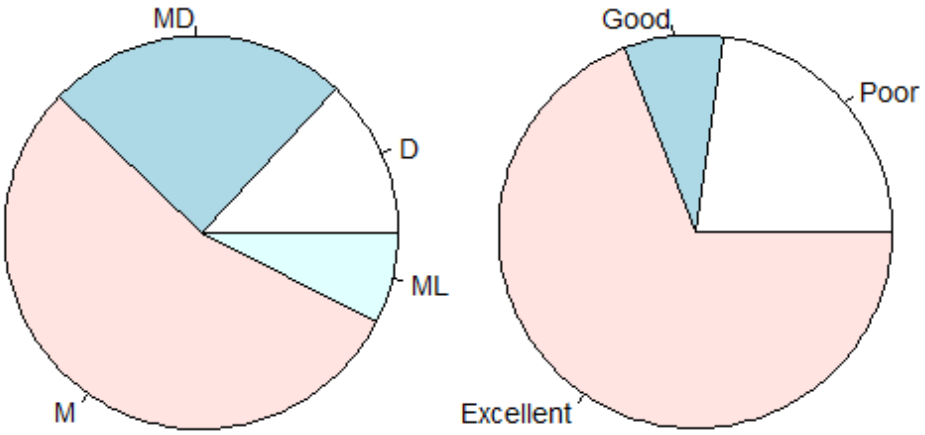
attempt, random forest uses an ensemble of multiple weak learners which often outperform a single decision tree.







Figures 4-6. Binned densities of observed values for numerical variables *sat*, *weight*, and *width*.



Figures 7-8. Category proportions of observed values categorical variables *color* and *spine*.

3. **Imputation and Results Analysis**

After 5 imputations across all variables in the dataset, applying 50 iterations per imputation, we plotted several plots to compared original observed values and imputed values. Most imputed values fall closely to the observed values, with the exception of *sat* (Figure 9) due to irregular distribution.

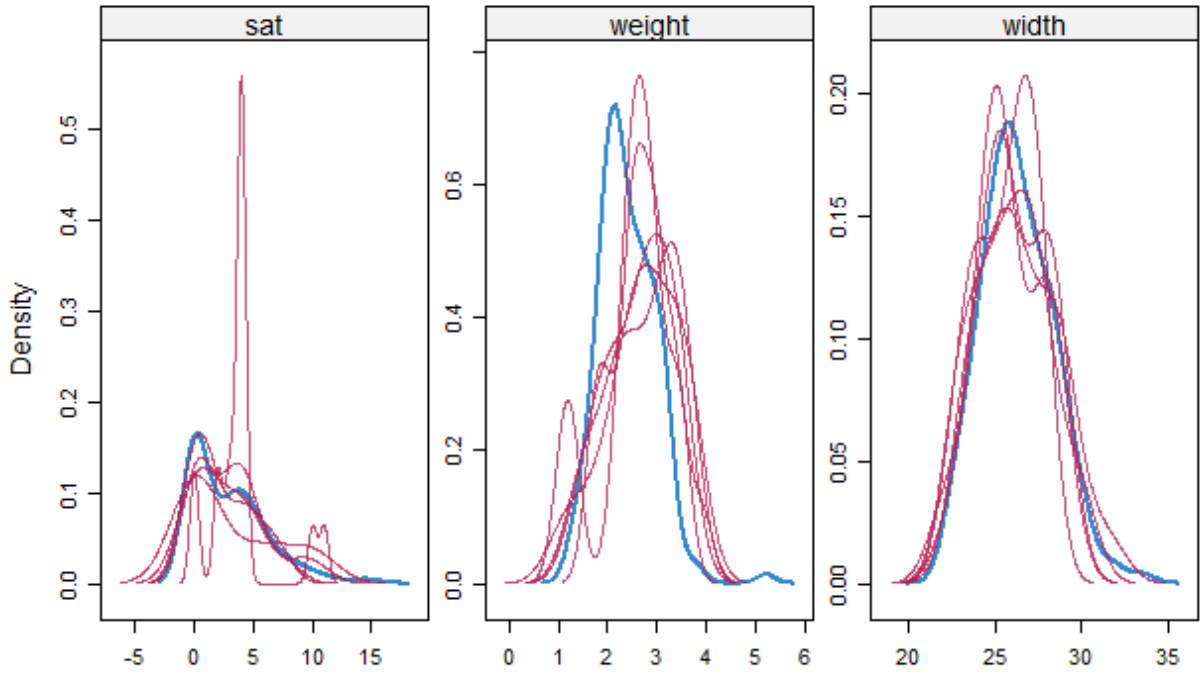


Figure 9. Density plot of observed values vs. imputed values for numerical variables.

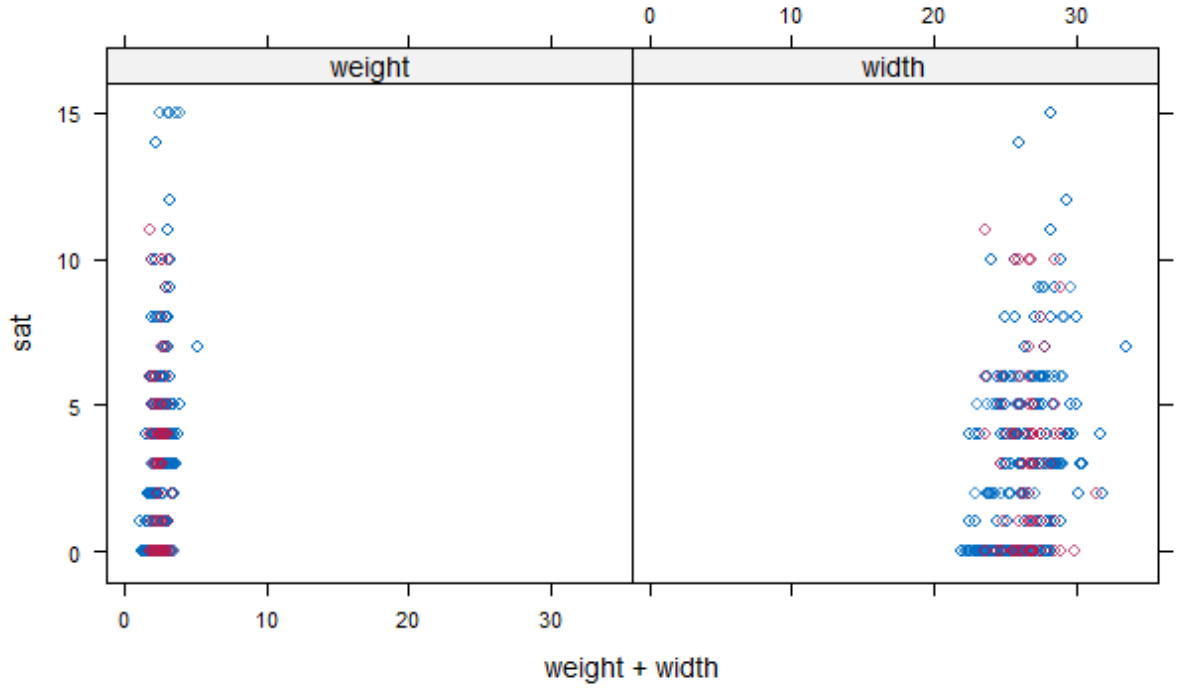


Figure 10. XY Plot of observed values vs. imputed values for numerical variables.

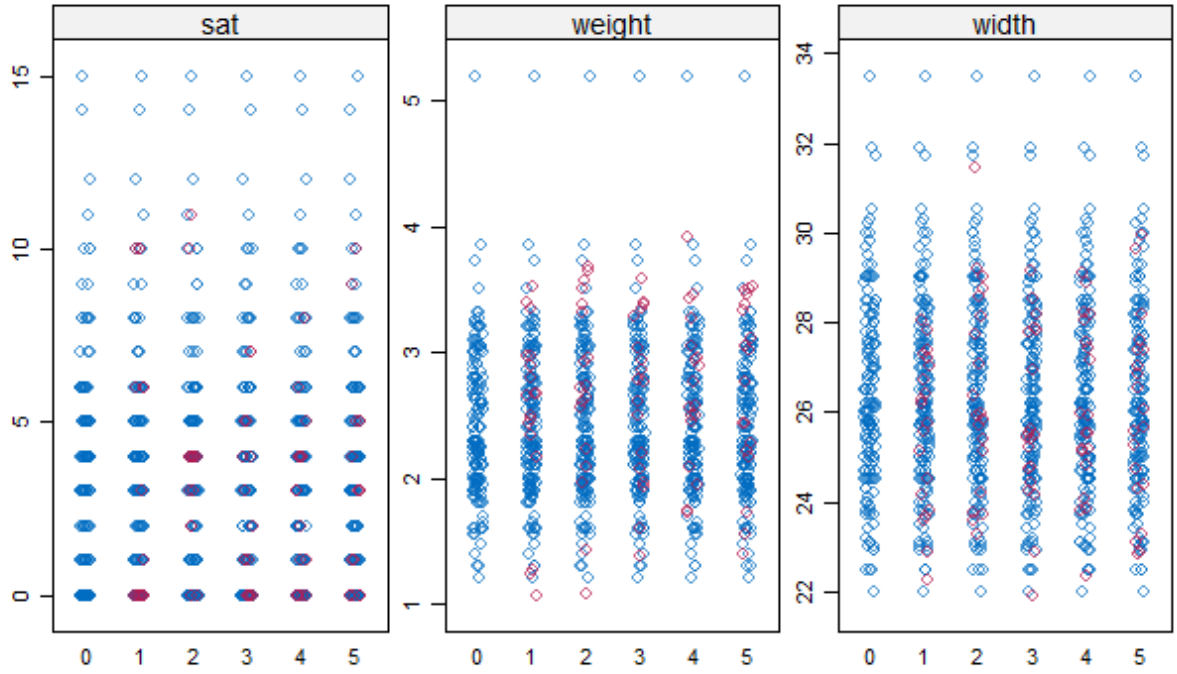


Figure 11. Strip plot of observed values vs. imputed values for numerical variables.

4. See code.

5. Conclusion

	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	1.7637613	4.6841376	0.3765392	163.0193	0.707005736
2	weight	2.3407034	0.8417303	2.7808234	163.0193	0.006060848
3	width	-0.1923317	0.2404057	-0.8000298	163.0193	0.424857792
4	colorMD	-0.3149026	0.8142897	-0.3867206	163.0193	0.699467362
5	colorM	0.6684225	0.7463705	0.8955639	163.0193	0.371806165
6	colorML	1.3290443	1.1392216	1.1666249	163.0193	0.245065654
7	spineGood	-0.7130178	0.9950262	-0.7165819	163.0193	0.474657467
8	spineExcellent	0.2844562	0.6262936	0.4541899	163.0193	0.650296285

Figure 12. Coefficient estimates.

In conclusion, we were able to implement imputation using the mice.impute.sample method. This method was chosen because the observed values for the numerical values are always non-negative and this method imputes a random sample from observed y data (in this case, satellites). Without having knowledge of what the true population distribution was shaped like, this was the most plausible way to compute our estimations. Based on our results, we found that *weight* is the only variable with a very small p-value, which makes its coefficient estimate statistically significant, while all other variables are essentially negligible when it comes to predicting the number of satellites.