For my data analysis report number 3 project, I would like to the predict the average avocado price based on the avocado's characteristics, locations, and size. The dataset I picked has a small sample size which would makes the training and testing data even smaller in sample size when we split up the dataset. In Gelman's video, he briefly mentioned about plotting and predicting with a small sample size data. This is leading to high variability of the prediction. In addition, I like the idea of sub-setting the data into train set to build the model and a hold-out test set to validate the model from the paper "Prediction, Estimation, and Attribution". For my data analysis report 3, I would split my data into train and test set using random partition. This way, it would make the prediction more random and the model would consider a random train set to make our prediction less biased.

Prediction, estimation, and attribution are good methods for statistical inference tools. We can validate the quality of the model using predictions and the empirical of response variable. With this being said, there are variety of criterion that we can consider when validating the models. We can use mean square errors of the empirical response variable minus the predicted values. This would give us the metric that would tell us the model that have least mean square error would be better since this model is predicting better and have less errors.

In my project, after generating prediction and attribution, I can be able to say that my final model for avocado price would be good or not. With the predictions of avocado prices, I can utilize the model that have least square error to predict the average avocado prices. With the more accurate model, I can predict average avocado prices more effectively using the predictors given in the dataset. In addition, the inferences would include the analysis of adjusted R square, AIC values, and the significant of test statistics of each predictors in the model. Therefore, predictions and attribution would ensure that my model is a good model or not.

Prediction and attribution compliment each other in many ways. To validate the model and report my findings to key scientific stakeholders, there would need to be more than one criterion that would prove my model to be a good model. Attribution would give the stakeholders more information about the training data itself. This will help the stakeholders knows more about their historical data. Furthermore, the predictions would prove to the stakeholders that the model is doing a good job in predicting future avocado prices or not. This will ensure that stakeholders would have an idea of how good our model in term of using historical data to predict new prices. In addition, checking the predicted data with the empirical avocado prices would also give the stakeholders a better idea of how small/large the errors are when using this model.