REPORT

- **Question 1:**

Fixing the ‘bathroom’ variable involved separating the integer and decimal values of the variable for each observation and multiplying the latter by five and adding it to the former (e.g., if a house had a reported value of 3.1 bathrooms, then, the corrected variable is $3 + 5 * .1 = 3.5$.

The ‘state’ variable was dropped since all observations are in California. Dummies were generated for the remaining covariates (‘city’, ‘maintain’, ‘sidewalk_access’, ‘kitchen_quality’, and ‘home_functionality’) and ‘price-range’ was factorized in order to train the machine to forecast it.

A Random Forest and Logistic Regression models are used since both are the better approach when dealing with a discrete categorical variable that needs to be forecasted. Regardless, the accuracy of the model appears to be sub-par, with the former hovering around 37% and the latter, 39%.

- **Question 2:**

Dropping the corrupted observations does increase the accuracy of both models but still the forecasting capabilities are still unsatisfactory, with accuracy scores for the Random Forest hovering around 38% and for the Logistic Regression, 40%.

Rather than using the suggestion by the colleague, we could build a machine to predict the number of bedrooms for the corrupted observations from the cropped data and then plug the forecasted values back into the main dataset. Losing observations is never a

good deal since we want randomness in our data but, simultaneously, adding variables that were forecasted can generate endogeneity problems.

- **Bonus Question:**

We fixed the data through the same steps used in the first question, but now, the model utilized is a simple Linear Regression model, since we have a continuous dependent variable to be predicted and the housing market is suitable for a Hedonic Pricing model. For this scenario, our model is actually satisfactory, with R^2 scores hovering around 81%.