Hello. In this Jupyter notebook, I take a dataset based on diamonds from Kaggle, clean and analyze the data, and then use various regressor models on it to see which model is the most accurate.

After loading the dataset, I first changed all the categorical columns to numerical columns by mapping each value to a number according to the hierarchy I described in my comments. This allows me to see the correlation coefficients of these columns. Then, I create a correlation matrix on the edited dataframe. As you can see, the columns carat, x, y, and z all have a huge correlation with price and each other, while the other columns have a low correlation with price. X is the length, Y is the width, and Z is the depth of the diamond.

I create some visualizations to show why some columns have such a low correlation with price. The scatterplots of depth vs price and table vs price show no linear relationship. Since the categorical columns have discrete values, I decide to show the interquartile range, a measurement of spread, of each possible value of the categorical columns. Besides 40% of the diamonds having an “Ideal” cut, no categorical value exceeds 25% of the diamonds. A lower spread is ideal for predicting the price. I also show the IQR for the middle 50% of values from the carat, x, y, and z columns. I create a function that gets the IQR of the price based on a given condition to save space. None of the IQRs for the middle 50% of the highly correlated numerical variables exceed $2418. All of the different cuts have a higher spread than $2418. 3% of the diamonds have a ‘Fair’ cut and yet it has such a high spread. All of the different colors have a higher spread than $2418 as well. 10% of the diamonds have the ‘I’ color and 5% have the ‘J’ color and yet their spreads are 140% higher than $2418. All but 2 of the clarity values have a higher spread than $2418. Overall, these visualizations show that cut, color, and clarity are terrible predictors of price. The columns I have left are carat, x, y, and z. 20 of the rows have a 0

either as the x, y, or z value so I remove those rows since they are clearly incorrect. I create some scatterplots to show the relationship to see if there are extreme outliers and to show the relationship between these columns and price. There are no extreme outliers for X. There are 2 Y values above 30 which are clear outliers, so I remove them. I remove the Z outliers as well. Carat has no extreme outliers.

Since x, y, and z, which are the length, width, and depth of the diamond are all correlated with each other, I decide to create a feature called volume which is the product of x, y, and z. Volume has a higher correlation with price than x, y, and z. I'll be using volume and carat as my two features in my machine learning models to predict price. I create train and test subsets and shuffle the data before splitting. I found 10 models in the sklearn library to try out. All of these models are regressors, need no custom parameters, unique from each other, and they all run. I put all of the models and their corresponding names in a list of tuples and create a dataframe to put my metrics in. Then I use a for loop where I cross-validate each model and get a list of metrics including root mean squared error, the r2 score, and mean absolute error. I put all of those metrics into a dataframe.

Finally, I create some visualizations that show the performance metrics of each model I used. The green bars represent ensemble models, and the blue bars are not ensemble models. The top 3 models for each metric are all ensemble models. For mean absolute error, the Passive Aggressive model has the highest and the Gradient Boosting model has the lowest. This is also true for the root mean squared error. Gradient Boosting also has the highest r2 score and Passive Aggressive has the lowest. This clearly means that the Gradient Boosting model is the best one while the Passive Aggressive one is the worst. I'll consider using Gradient Boosting for future machine learning projects. Thank you.