SUPERVISED MACHINE LEARNING

DATA ANALYSIS PROCESS:

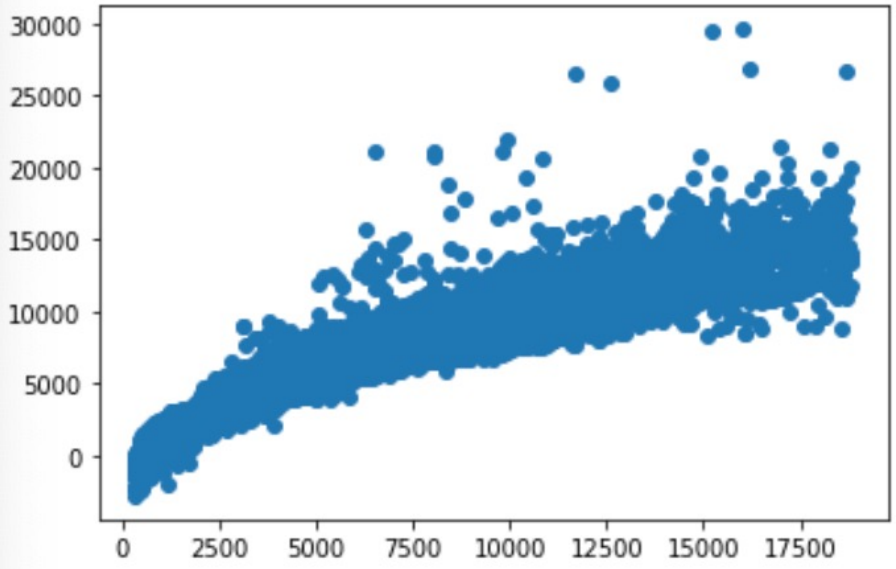
(DATA WRANGLING)

1. To begin the process of analyzing the diamonds dataset, I first imported all the packages required for each step of the project, namely packages through ‘sklearn’, ‘numpy’, ‘matplotlib’, and ‘pandas’.
2. Then, I imported the dataset, and reviewed it. Three out of the four independent variables needed to be converted into numeric variables:
    - First, created a dictionary for each of the categorical variables (ie, cut, color, clarity)
    - Second, mapped into the ‘diamonds’ dataset
3. Lastly, I defined my ‘x’ and ‘y’ variables in order to begin the process of supervised machine learning.
    - ‘x’ = carat, cut, color, clarity
    - ‘y’ = price

DATA ANALYSIS PROCESS:

(SUPERVISED MACHINE LEARNING)

- I decided to do the Train-Test Split with a 60/40 split: using 60% of the data for training and 40% of the data for testing.
- Next, I needed a linear regression model in order to interpret the supervised machine learning model accuracy
- This is the plot I received using my predictions:
- The accuracy looks good on the plot and printing the "Score" of the testing, we get back that the model is accurate approximately 90.4% of the time.



- Accuracy looks good!

DATA ANALYSIS PROCESS:

(EXAMINING ERRORS)

1. I decided to conduct another way to quantify the residuals of the data by examining the errors using the *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, and *Root Mean Squared Error (RMSE)*
2. Each mathematical way to examine error returned acceptable numbers, considering they are judged between zero and infinity.
3. To ensure that bias hasn't been introduced to my model through the split method, I concluded my project by conducting k-fold cross validation. I split the data into 6 different datasets (3 sets for training, 3 sets for tests), then my model applied 3 training-test data sets which it then accuracy checked for 3 test-training data sets.
4. **All trained models are accurate around 90% of the time—all models were nearly the same as the previous testing—meaning the model fits very well with the data.

CONCLUSION:

THIS MODEL FITS WELL WITH THE DATA AND WILL BE

ABLE TO PREDICT THE PRICE OF DIAMONDS ACCURATELY

ABOUT 90% OF THE TIME.