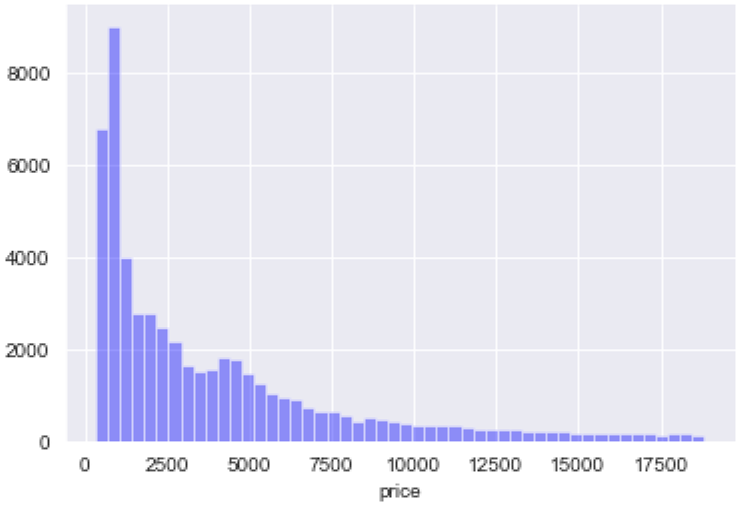
# Diamond Pricing System Final Report

## Business aligned goal

The hypothesis I formed in the problem statement for this report was “What opportunities exist for interested parties (sellers and buyers/consumers) of diamonds to predict the value of a particular diamond using 10 features better than the natural model from historical prices with at least 75% accuracy, based on observations found from the Diamond Price Dataset via Kaggle.com.” To answer that question I performed a detailed data analysis while building a model for deployment.

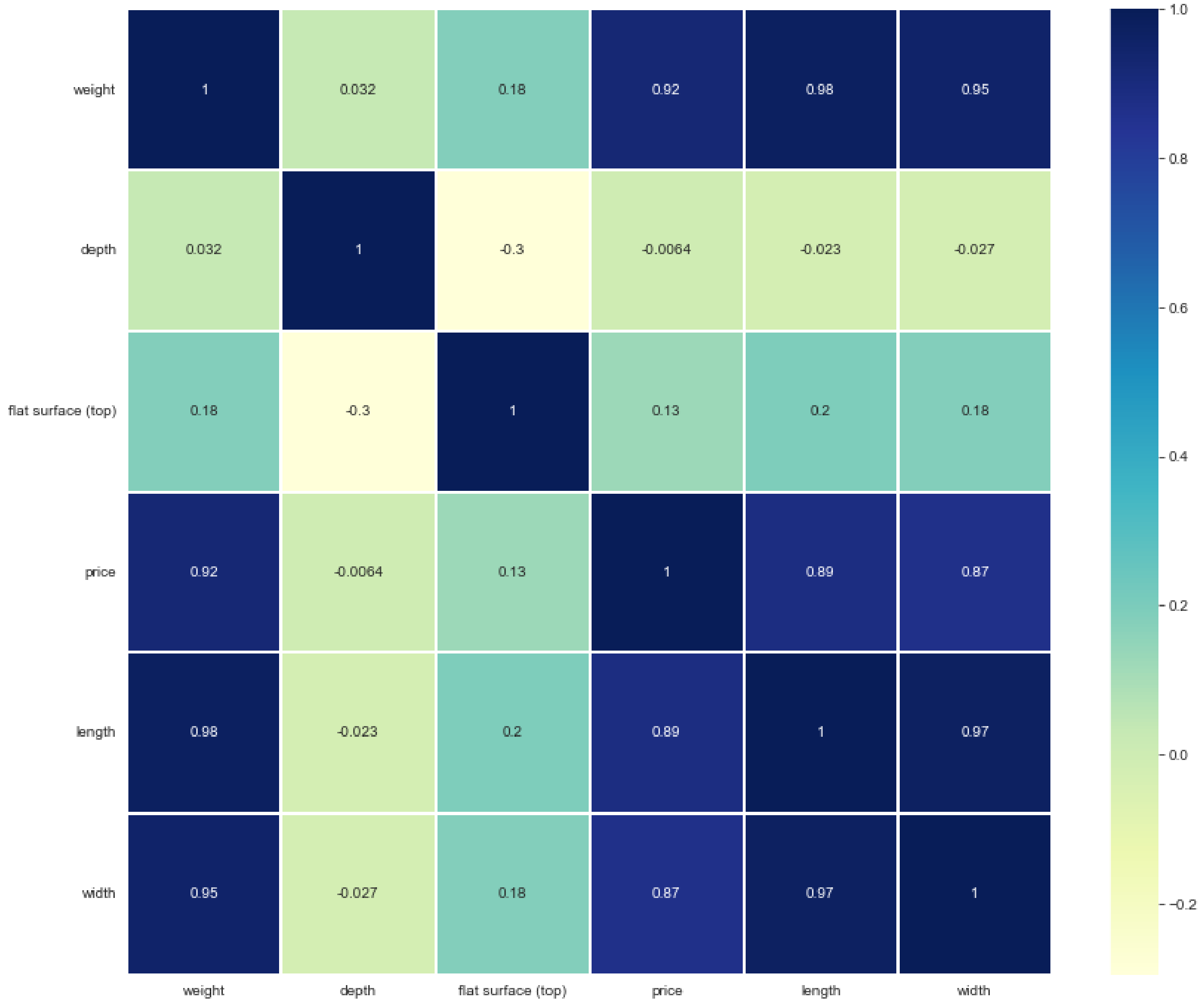
## Summarized Findings

Initially when exploring the data, we could see that the lowest prices start at 326.00 in US dollars with 2 observations of that amount and can raise all the way up to 18823.00 in US dollars with 1 observation at that amount.



# Diamond Pricing System Final Report

From the histogram that was plotted, we can see data points for the price variable are skewed to the right and are of an exponential distribution. This makes sense as diamonds are a luxery item and the more expensive they are the less of them are usually kept on hand. Essentially, the bulk of the group should be toward the cheaper prices as it is cheaper to keep those type of diamonds in bulk. However, this was a problem for a regression model if not managed effectively.



# Diamond Pricing System Final Report

A glance at the heatmap above shows how strongly correlated the different diamond measurements are with each other.

- In particular the strongest correlations are varibles width with length as 0.97 and length with weight as 0.98.
- If we are planning to treat all variables as independent in our modeling process, co-linearity rules would be violated.
- A model technique like Random Forest or a decision tree might be a better fit for the data.
- Random Forest or decision tree models would not negatively be impacted by high variable correlations.

## Model

When determining the price for model for diamonds several models were used. With the first model there is a visible positive correlation, as the model has not been totally unsuccesful, but it's clear that it is not maximally accurate: diamonds with an actual price of just over 11250 have been predicted as having prices from about -4000 to about 6,000. The overall score of the model was approximately 92%. However the best model out of all four used was the Lasso Regression model. Given there are many features of diamonds that we would want to minimize for simplicity, this is a good model choice. The overall score of the model is approximately 90% but its added complexity is a good reason to choose it over the better scoring but simple, first model.

# Diamond Pricing System Final Report

The defining assumption for the model is that other current practices are benchmarks for the market (or diamond-buying public) that sets the scope for the model. There are also other constrictions/assumptions that could be considered or that the model does not assume like fair market practices and actual pricing of value for each individual diamond. The model shows the expected number of observations or diamonds of over the 13,350 and, on average. The provided data includes the original variable and features of the raw data.

## Recommendations

Based on aligned business goals previously established regarding the following questions:

- What diamond features exist and are considered important based on industry standards and what data can be used to determine their value other than price alone.
- Why are some diamonds more valuable than others?

a few options are recommended. There can be many use cases for this type of prediction especially for an audience of either the seller, buyer, or consumer and majority industry scenarios. Since diamond price is determined by a few set of parameters the price can predictable and as a result standardized. The success criteria for this case-study is based on the ability to predict the value of diamonds at 75% accuracy resulting in a more reliable model to predict the valuation of diamonds more efficiently than historical prices without the model. This model represents competitors and their rationale for a given diamond price while creating benchmarks to not only measure their current pricing but predict future pricing.

*Three potential scenarios that can be represented using our model is the following:

1. Focusing on what diamonds are more valuable than others and being able to plan for those higher ticket items in the future.

2. Using the current prices for training purposes. Adjusting or changing that model to match factors or predictors is a critical tool that can be used in the future by any audience interested in understanding diamond prices better.

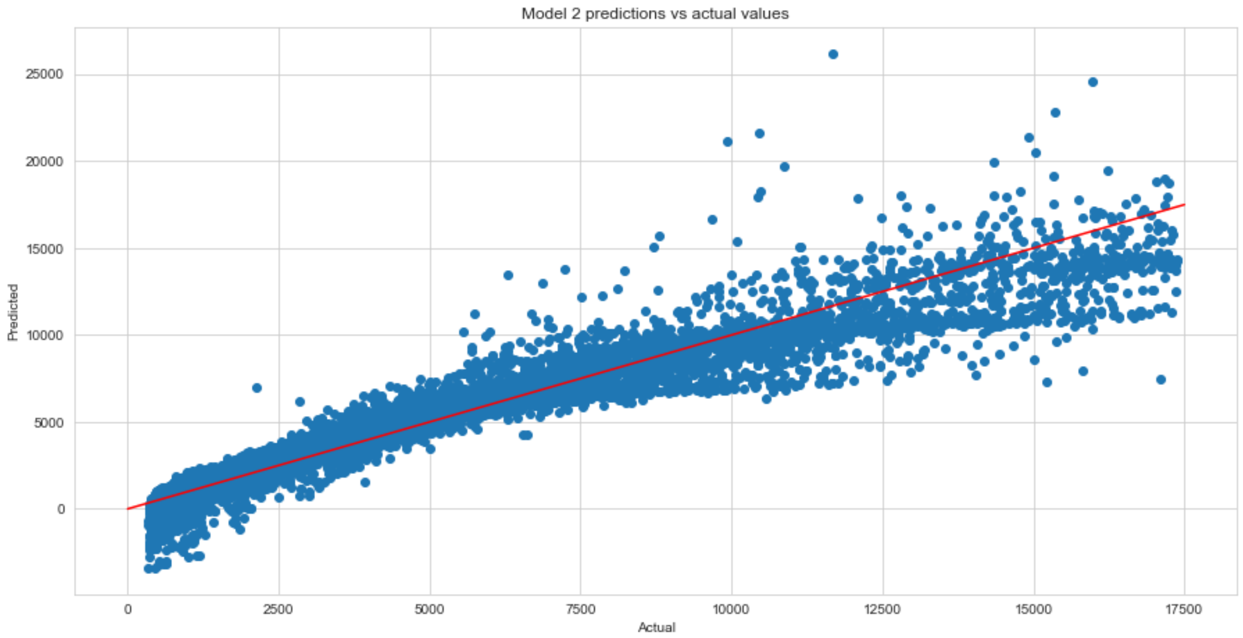
3. Selling highlighted features for lower priced diamonds and basing sales plans around that concept.

## Potential deployment and maintenance

For future improvements, I recommend running a linear regression model because it is simple but also leads to similar reliability as Random Forests models when it comes to this data.

Testing can also be used by attempting to use FIFO or First In First Out historical data always to observe how accurate model results are to actual results to avoid potential valuation loss.

# Diamond Pricing System Final Report


Model 2 predictions vs actual values
Predicted
Actual